Open Research Online

The Open University's repository of research publications and other research outputs

Authenticity and the assessment of modern foreign language learning: The problems of designing authentic tasks and devising and applying criteria for moderated assessment and evaluation in the examinations of the International Baccalaureate for Modern Foreign Languages

Thesis

How to cite:

Israel, John Black (2003). Authenticity and the assessment of modern foreign language learning: The problems of designing authentic tasks and devising and applying criteria for moderated assessment and evaluation in the examinations of the International Baccalaureate for Modern Foreign Languages. EdD thesis The Open University.

For guidance on citations see FAQs.

© 2003 The Author

Version: Version of Record

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

UNRESTRICTED

Name: John Black ISRAEL
Personal Reference Number: M 7259477

AUTHENTICITY AND THE ASSESSMENT OF MODERN FOREIGN LANGUAGE LEARNING

***The problems of designing authentic tasks
and devising and applying criteria
for moderated assessment and evaluation
in the examinations of the International Baccalaureate
for Modern Foreign Languages***

Doctor of Education (Ed. D.)

18th. December 2003

CONTENTS

PART I

INTRODUCTION:

THE AIMS OF THE RESEARCH AND ITS CONTEXT

PART II

THEORETICAL BACKGROUND

PART III

RESEARCH METHODS AND FINDINGS

PART IV

CONCLUSIONS

APPENDICES

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Notions of authenticity often determine aims in communicative language teaching and learning. This research describes and develops theories of authenticity in assessing and evaluating such activity. Concepts are defined for mapping and exploring the **International Baccalaureate Organisation**'s *Diploma Programme* for *Group 2 Languages.*

The empirical focus is *Language B*, *Standard Level*, a programme for intermediate foreign-language learners. Attention is paid to formal assessment in listening and speaking, reading and writing in French. It includes the delineation of boundaries, investigation of rubrics, design of tasks and their standardisation, language use in criterion-referenced assessment, with the moderation and evaluation of results by grades.

In measuring performance, 'target' language communication is investigated, insofar as definable and assessable through reference to authenticity. Commonly-used theoretical and practical categorisations emerge as subjective, imprecise and contestable.

Three methods are employed to identify, describe and understand the programme, together with the language use it entails. They provide complementary perspectives for conceptualising authenticity.

First, samples of **IBO** documentation are analysed for illuminating theory. Understandings are developed and refined through observation of the programme in practice.

Alternately, constraints on learner participation in assessed language-production for authentic communication are examined. Influential in any situation, they appear particularised in 'high-stakes' evaluation.

Understandings are also derived from analysis of qualitative and quantitative data, sampled from a range of formal assessment sessions

and manipulated experimentally. Responses to specific tasks are scrutinised.

Through developing criteria for identifying, analysing and evaluating language-based authenticity from this data, the research seeks:

- to assess validity in devising standardised tasks for authentic language use within set rubrics;
- reliably to correlate qualitative, criterion-referenced assessments with quantitative evaluation;
- to determine regularity in grading significant qualities of formally-assessed language;
- better to understand authenticity as a concept for guiding these aims;
- to identify theory and practice that distinguish the programme researched as a view of pedagogy and learning, through investigation of its products.

The research offers description, analysis and critique of programme planning, administration and outcomes. Its conclusions indicate authenticity as conceptually viable for assessing language use. Without decreasing reliability, construct validity may be enhanced. Anomalies previously found 'difficult to assess', are reduced in incidence and the fit improved between programme philosophy and practice.

Through measuring task-based language for authenticity in determinate settings, evaluation verdicts may be more consistently and explicitly justified, enhancing the potential credibility of the given programme amongst its users.

ACKNOWLEDGEMENTS

Many have been of invaluable assistance throughout the course of research for this project, and I should like to thank them, by recording their names and some reference to their contributions here.

First and foremost, my deep gratitude goes to David Scott, Professor of Education at the University of Lincoln and my supervisor at the **Open University**. His advice was inspirational, hard-headed, extensive, always willingly and generously offered with professional concern and unflagging enthusiasm that invariably encouraged confidence to persevere throughout.

No less gratitude is due to John and Diana Graves for their munificent hospitality and patronage, their listening ears and their morale-boosting discussion, freely offered at regular intervals throughout the development of my work.

At the **International Baccalaureate Organisation** my thanks are offered to many. Without their patient acceptance of the intrusiveness of research and open-minded support, it would not have been possible to proceed. That the encouragement they gave was so regularly generous in time, effort and interest was a much-appreciated bonus. George Pook, Anne Scott and David Ripley have been very notable in this respect. Equally, much gratitude must be extended to Helen Drennen, John Shillaw, Roger Brown, Christine Trumper, Renėe Joly, Helga Eckart, Peter Kino, Jane Wade and others, especially those in the committees and moderation groups associated with *Group 2 Languages*, as well as the teacher 'guinea pigs' in attendance at particular **IBO** programme-training workshops in Edmonton and Halifax, Canada, London, Malta and Denver, Colorado, led by myself. With all the willing engagement in discussion that took place, I could attain far more in the company of fellow travellers, than I would have on my own.

I should also like to recognise and acknowledge with thanks, both staff and fellow students of the *Ed. D.* programme at the **Open University**, Martin Jephcote of the School of Social Sciences, University of Cardiff, Val Klenowski of the Institute of Education, University of London, and Chris Woodward of the University of North London, for their professional and personal interest and advice.

At the **Istanbul International Community School**, thanks are due to professional and personal, 'critical friends': Regina Alfonso, Michelle Altuğ, Jill Bamforth, Marie-Christine Monaco. For document processing, generous assistance was always freely given by Lane Ceylan, John Dewsnip, Jennifer Gökmen, Sezai Kara, Ömer Kipmen and Cemal Ulu.

I am also indebted to the libraries and staff of Bristol University Education Department, Cardiff University Education Department, University of London, Institute of Education, Open University Education Library, The British Council Library, Istanbul, Turkey, for their handling of my enquiries and guidance in accessing their collections.

Other critical friends and facilitators have also been significant in supporting me throughout the progress of the research. I should like to conclude with particular thanks in recognition to my sisters, Jocelyn

Stewart and Kate Israel, and friends of very long standing, Nick and Lizzie Gould and Martin and Lisa Jephcote.

PART I

INTRODUCTION:

THE AIMS OF THE RESEARCH AND ITS CONTEXT

CHAPTER ONE

HYPOTHESES AND KEY QUESTIONS

Initial Approaches and Rationales

In commencing research, two general hypotheses were postulated. Conceptualising authenticity could illuminate understanding of communicative processes and their products. In assessment contexts, such concepts could serve analysis of authentic language use. The combination of theory with scrutiny of situated practice provides a rationale, point of departure, direction and scope for this investigation.

First explored through pilot research[1], the hypotheses were refined within two distinct perspectives: one derived from theoretical tradition, the other practical and emergent from assessments of non-native, second-language performances, produced either in international school classrooms, or by international students for evaluations under the *Diploma Programme* of the **International Baccalaureate Organisation**[2]. An established philosophy of authenticity was related to empirically-developed theorisations of communicative usage. In varying settings, day-to-day, monolingual and 'target' language interactions between the researcher as teacher, assessor, **IBO** *Examiner* and *Moderator*, and multilingual students of differing linguistic backgrounds, led to reflection, re-questionings of ideas and deeper comprehension. The twin perspectives are complementary, integrating theory and practice within a single research-design.

Through analysing initial evidence, the ontological, epistemological and ethical dimensions of foundational propositions were investigated and recounted as explicitly and consistently as possible. Progressive focussing on salient features, both theoretical and empirical, indicated bounds for a formal project. Interlinking research questions were derived from this experience, delimiting scope and determining the detail of the research design.

Identifying ontological features of authenticity provides foundations for explaining phenomena traditionally categorised as structural, behavioural, psycholinguistic or purely linguistic, and discernable in stable, recorded evidence of language use. Positivistic approaches to pedagogy, learning and assessment have stressed atomisations of language, rote-memorisation of model structures and vocabulary, unmodified, inflexible practice through repetition, with discrete, point-in-time testing of standardised linguistic knowledge in units, measured by matching items of candidate-response to assumedly incontestable, authoritative and reliable norms. These are often decontextualised, unchanging and officially-sanctioned. For 'objectivity' in measuring language quality, reliability becomes of greater concern than construct validity. The promotion and assessment of communicative and interactive, socio-linguistically contextualised performance skill is often of secondary importance. Intersubjective, interpretative approaches to evaluating quality in communication between two or more partners interacting through use of common language, has been eschewed as insufficiently rigorous, difficult to replicate and overly-restricted by the contingent particularities of unique assessment situations in which language users perform.

Whilst addressing the ultimate constructs and purposes of language teaching and learning, performance-based approaches to assessment, measuring quality through criterion-referencing to idealised descriptions, rely fundamentally upon literally subjective construal for establishing validity. Communicative interaction between individual speakers and listeners, readers and writers is evaluated. Recourse to repeated moderations, multiplying interpretations in order to arrive at consensus, rather than strictly 'objective' point-scoring for 'correct' answering in a positivist sense, does not remove subjectivism from this assessment process. Recorded performances in unique situations, may provide evidence of knowledge and skill, but cannot be easily and rationally quantified. For positivists, criterion-referenced, performance-based measurements of language quality are all too frequently, unacceptably unreliable.

Over the last half-century however, interest has grown in devising alternative forms of language pedagogy, learning and assessment. In particular, concern has been expressed that positivistic and behavioural approaches to acquisition and use unduly restrict the development of 'communicative competence' (Hymes, 1971) and hamper 'authentic' performance. They rely on a predictable, precise and thus 'unrealistic' replicability of task, requiring responses that ignore the particularities and almost infinite variability of contingent, temporal and socio-cultural contexts. This hinders the development of learner qualities such as spontaneity, fluency, adaptability to circumstance and appropriateness of usage. For all settings in which performance is required, concern for practical goals in modern, foreign and second language-learning has led to the growth of interest in 'authenticity', in defining its meaning for language use and in searching for valid, reliable and practicable means for assessing and evaluating its quality in productive performances.

The **IBO**, through its explicitly international philosophy, aims and objectives for language development, has historically been at the forefront in shifting educational paradigms, given its international orientation and consequent stressing of central importance for bi- and multilingualism in its *Diploma Programme.*

However, 'authentic language use' is often inadequately described and too loosely categorised for rigorous assessment, despite widespread recognition of **IBO** criterion-referenced designs for evaluating second and foreign-language performance. Throughout the research, empirical evidence of authentic production has been identified, described and analysed with continually-increasing refinement, clearly demarcating conceptual components for use as initial benchmarks and signposts in exploring programmes such as these. Componential categorisation of authenticity permits experimental development of descriptive criteria for assessing samples of language-production as valid, or deficient exemplars of authentic communication. When applied and contrasted with criteria from existing schemes, supplementary vantage-points for data-analysis and interpretation are created. Unity in research-design is founded on a clearly-delineated, common body of empirical evidence, derived from situated language-productions within a single programme and accumulated from repetitions over seven years of identical assessment sessions.

From the outset, experimental criteria for identifying and measuring relevant components of language use were recognised as ideal representations of authenticity in communicative relations, being derived from theory, whether 'espoused' or 'in practice'[3]. From triangulating theoretical with empirical viewpoints, experimental categorisations improve understanding of what constitutes authentic expression. They are tools for research, acting as comparators for analysing alternative systems. Within the *Diploma Programme*, critical appreciation of policy and practice is developed in this light[4]. The approach assesses the consistency of distinct epistemologies (notably typifying psycholinguistic, sociolinguistic and communicative analyses of language-production, within their characteristic systems of measurement). Even when implicit, such disciplines provide well-known means for explaining and evaluating language use. Promoting 'authenticity', they colour **IBO** discourse, outlining conceptualisations, aims, objectives, course-descriptions, requirements, assessment, moderation and evaluation rubrics, as well as associated procedures for putting theory into practice.

To summarise, the **IBO** characterises language-performance as evidence for acquired skill and knowledge, integrating psycholinguistic, sociolinguistic (and occasionally aesthetic) dimensions with structural features of 'pure' language. In a commonly-defined and accepted mode, it should 'communicatively' link speaker with listener and writer with reader. Phenomenological conceptualisations of authenticity as relationships between co-producers and receivers of communication may be inferred from any linguistic production[5]. The selections and emphases of 'facilitators' and performers, guided by such discourse, influence the teaching and learning of *Group 2 Languages* and shape the evaluation of resultant usage[6]. Theory is evident in statements of overall, curricular 'philosophy' and in definitions of boundaries for programmes, arranged in coherently-graduated, hierarchical sets. These are published with workable schemes for internal and external assessment, and external evaluation. All mould the language use studied, and are researchable through applications in organisational practice[7].

Investigating situated, communicative language-performance facilitates understanding of matchings, or mismatchings, in embedded theories of pedagogical, curricular and assessment knowledge, with an implicit ontology unifying the **IBO**'s epistemological outlook. The organisation's statements render authenticity significant, linking an 'espoused theory' of philosophy and aims for particular programmes with judgements of value emerging from all assessments. Axiological effects inhere in practice, and are made explicit in published evaluations of performance.

For French as a second language, specifications, analyses and classifications of positionings[8], understandings and communications of particular individuals are also investigated. Explorations of authentic expression may thus be independently assessed and evaluated by readers familiar with, and experienced in processing the primary sources of data, provided by candidate productions. **IBO** working practices were observed and analysed, and relevant documentation scrutinised, with evidence for critical analysis selected from a comprehensive range, including: samples of records from the organisation's administrative archive; observations of moderations and evaluations; and sets of assessments, as exemplars of language produced orally for *Internal Assessment* and in writing for external examination.

Exploring authenticity through complementary perspectives, one theoretical and 'literary', the other practical and 'grounded', facilitates comparison. Congruencies, similarities and differences are analysed, classified and discussed. Such data-interpretation serves to enrich understanding of any programme for measuring second-language knowledge and performance, devised by alternative organisations for similar levels of competence.

The Origins of the Hypotheses

Dual perspectives for conceptualising authenticity emerged from selections of literature, both theoretical and practical in focus, and the personal and professional interests of the researcher[9]. Particularly relevant too, is experience in parallel employment as a teacher of French as second or third languages, to students aged 11 to 19, in non-selective, multicultural and multilingual, though English-medium, international schools[10], and in various roles, as an **IBO** *Examiner, Moderator, Reviewer*, *Teacher-Trainer* and *Teacher-Observer*[11]. From such experience, preliminary, *a priori* conceptualisations of authenticity were developed[12]. Derived from pilot work, initial statements represented researcher understandings on commencing research. Concomitantly, notions were inferred from **IBO** usage, embedded within its discretely-categorised, yet interlinking language programmes. Additional understandings arose from regular, though informal communications and exchanges throughout concurrent employment[13].

Subsequently, authenticity emerged as a general criterion guiding specifications, selections, and creations of *content* in given curricula for teaching, learning, language-production and assessment. It also related to processes of choice, comprehension, completion, assessment, moderation and evaluation of communicative performance in *tasks* through which this content is expressed. The linkage of *Message* with *Task*, in examples of individual performance recorded at distinct points in time and under specific administrations of set assessments, further unifies the perspectives identified for study.

A Preliminary Understanding of Authenticity

Theories of authenticity represent sophisticated, dialectical constructions of linguistically-based and culturally-integrated, valued knowledge, both intrapersonal and interpersonal, through reflexive engagements over time of 'self' with 'self', and skilful interactive engagements of 'self' with 'other'. Authenticity is a psychological feature of relationships, formed through personal, individual history, linguistic features of commonly-understood language and sociological aspects of temporally-particular, socio-political and socio-cultural worlds of communication.

Under **IBO** assessment and evaluation for exotelic purposes of candidate certification[14], this personal, socio-political and socio-cultural world of language use is partially conditioned by needs to satisfy predefined criteria, employed in shaping, assessing, sampling, moderating and evaluating linguistic productions of candidates choosing to be examined. The **IBO** assumes responsibility in designing assessment systems and evaluating performances for external accreditation. Language is produced not merely for its own sake, but for the award of diplomas and certificates of attainment. Evidently, outcomes influence not only selections of courses and institutions for further learning at higher levels, but also the selection of applicants for admission to tertiary institutions of education. In the interpersonal and social relationships created, questions of balance and symmetry typify issues of 'power' in language-based expression (in its broadest political and cultural senses). Likewise, institutional 'power' partially shapes linguistic understandings and forms, produced in authentically-coherent, situated communication.

For research, authenticity has been defined as a working concept, enhancing construct validity and applicable to all stages of specifying and analysing the design, standardisation, assessment, sampling, moderation and evaluation processes of the **IBO**'s *Group 2, Language B* programme. It partially governs the design, standardisation and production of programme guidelines, rubrics and examination papers by the **International Baccalaureate Curriculum and Assessment Centre**[15] for internal and external assessments. It illustrates descriptions of situated language produced in engagement with such processes, both by **IBCA** personnel, teachers and teacher-assessors, and by candidates for assessment. Requiring repeated, interpretative moderations as the mode of establishing credible assessment reliability, questions of authenticity influence ultimate evaluations. Within this setting, the concept regulates data-analysis of:

- candidate understandings, choices of assessment task and responses;
- teacher understandings, selections, structurings and actions as interlocutors and facilitators of performance, particularly under *Internal Assessment*; published criteria, descriptors and procedures for assessment, as applied by **IBO** *Assessors, Examiners* and *Moderators* in measuring the qualities of candidate language-productions;
- ultimate, aggregated representations of competence and quality as diploma or certificate evaluations, graded on a seven-point scale of attainment; on occasion, the arbitration process for special cases[16], and appeals against outcomes from the foregoing procedures;
- published *Subject Reports*, with tables of equivalences relating qualitatively-sourced criterion-scorings to quantitatively-

enumerated evaluations for each certification session, together with comment and recommendations for future practice, provided by *Chief Examiners* for teachers of the relevant programme, amongst others[17].

In short, the empirical research-design has been determined, *a priori,* by the framework, perspectives and operational procedures of specific **IBO** programmes. All data-collection and manipulation has taken place within this framework. Through 'grounded' analysis of sampled language-productions, signs of inconsistency and hypothetical incoherence, progressively emerging as evidence from oral, *Internal Assessments* and written, external examinations, have been identified and discussed[18].

The Aims of the Research

IBO understandings of authenticity appear to derive eclectically from familiar traditions in describing and analysing language use, reported in specialist literature. These understandings have been refracted through the professional experience of successive generations of administrators, examination-designers, standardisers, teachers, teacher-assessors, examiners, moderators and evaluators, including the human experience of interacting with candidates as interested interlocutors and engaged readers. In any context, they concern interpretative theorisations, varying through the interactivity of communicative approaches to teaching, learning, assessment and evaluation, and under continually-evolving discussion. Hence an important research purpose has been to develop better understandings of authenticity in itself. Concomitantly, methods for more soundly securing interpretations of authentically-communicative language quality by assessors, moderators, evaluators and the researcher himself have been designed and tested. Thus, the significance of the concept for the evaluation of linguistic performances in purposive, social, culturally-specific, politically-charged and institutionally-constrained contexts, as well as for academic research, should subsequently become clear.

In summary, general aims concern:

- validity in devising standardised tasks requiring authentic language use within set rubrics;
- reliable correlation of qualitative, criterion-referenced assessments with quantitative evaluation;
- procedural reliability in determining the qualities of assessable language-production;
- better understanding of authenticity as an all-embracing concept, guiding the above;
- identification of theory and practice that integrate the *Diploma Programme* as pedagogy and learning, with assessment and evaluation of its products.

Key questions for investigation and resolution through research are derived from these aims.

The Definition of Key Questions

Preliminary experiences, their specific location within **IBO** frameworks, the point of departure, pilot results and general research aims, together generated four questions[19]. They are:

- What understanding of 'authenticity' and of related concepts, emerges from analysis of **IBO** publications and selected documentation, produced for internal administrative purposes?
- To what use is this notion put by the organisation in its task-design, standardisation, assessment, moderation and evaluation procedures, for both the oral and written production of the target language for **IBO** *Group 2 Languages* at a given level?
- What grounded understanding of authenticity in its various guises, emerges from discrete analysis of the work of selected examination-designers, standardisers, internal and external-assessment candidates, examiners, examination-moderators and evaluators?
- What inconsistencies in comprehension and practice can be identified through comparing the varying definitions, understandings and usages, both explicit and implicit, as outlined above?

Refinement and Development

For greater analytical scope, a triangulating experiment was devised to provide more than one vantage-point for interpreting samples of common data, albeit within the perspective of a sole researcher. Criteria defining features of communicative and authentic expression were developed from existing theory and integrated within a measurement system, respecting the moderated evaluation parameters of the **IBO**. Comparative analysis could better account for pilot-research evidence, and a preliminary framework be designed for organising less-focussed answers to the initial questions. Early indications of internal coherence were thus investigated, albeit tentatively, since data-collection, both qualitative and quantitative, depended upon interpretation, measurement and evaluation by the single researcher[20].

Subsidiary Questions

Notwithstanding this original approach, supplementary questions progressively emerged from research and were also considered. They led to further exploration of design constraints within the chosen **IBO** assessment programme, influencing forms and content for situated, authentic expression and requiring investigation. Subsidiary questions were formulated thus:

- In what ways and with what effects do the assessments studied 'position' the following:
    - the institution whose documentation is analysed?
    - the selected candidates whose work is analysed?
    - the moderator and examiner whose assessments and reports are analysed?
- What problems may be identified in candidate language-productions, attributable to institutional and organisational inconsistencies sourced at the **IBO**?
- What implications do any identified problems and inconsistencies have for applying assessment procedures?

Attempts to answer these supplementary questions better locate the project within an explicitly definable, socio-cultural and temporal context, allowing greater validity and sounder generalisation in conclusion. Improved comprehension of relevant issues of contextualisation may also guide subsequent developments, as initially envisioned.

CHAPTER TWO

THE ORGANISATIONAL CONTEXT OF THE RESEARCH

Preface

Much of this chapter is factual, with descriptions sourced from selected **IBO** publications, mainly for *Group 2, Language B* programmes[21]. For those unfamiliar with the organisation, other **IBO** references further contextualise investigations. Critical analysis of this supplementary evidence is not central to research purposes, though it is included where necessary for clarification, or for situating analysis and comment.

The International Baccalaureate Organisation

Since 1924[22], the **IBO** has worked to propose "a common curriculum and university-entry credential for geographically-mobile students"[23]. An early aim advocates developing international perspectives through promoting "intercultural understanding and acceptance of others by young people". Throughout the organisation's history, it has stressed the fundamental importance of international awareness in education, enabling students knowledgeably to compare their own society and selves, with others. For a *Diploma,* the **IBO** therefore requires study of more than one language to communicative purpose, with production of defined, minimum standards[24]. A comprehensive 'Baccalaureate' serves for assessing and evaluating internationalised curricula, administrable anywhere, and everywhere recognised as rigorous, valid and discriminating for transitions to tertiary level education.

In securing an international perspective, the **IBO** offers a wide range of languages, with assessment designed and evaluated by world-wide teams of experts[25]. The majority are *Assistant Examiners*, and many, practising teachers in **IBO** schools[26]. These teams are organised by, and responsible to an international team of *Chief Examiners*[27] distributed across all subject domains[28]. The latter are mostly selected from universities and colleges, specifically to represent many differing cultures, first-languages and academic traditions[29]. Together with recognised subject-specialists, they design and standardise assessments, examinations and evaluations, including the appropriate criteria for measuring quality. In each component, they also supervise the work of *Assistant Examiners* and *Team Leaders*.

The IBO *Diploma Programme*[30]

This is defined as a:

> "rigorous[31], pre-university course of study that leads to examinations; [.....] designed for highly-motivated secondary school students aged 16 to 19, [and] giving **IB** diploma holders access to the world's leading universities."[32]

It offers frameworks for a:

> "comprehensive, two-year, international curriculum [.....] that generally allows students to fulfil the requirements of their national or state education systems."[33]

Claiming to "incorporate the best elements of national systems", the organisation is "international" in outlook, insofar as its work is not based in any single, national design or set of values. It favours internationalism through integrating differing teaching and learning styles under a unified, criterion-referenced assessment scheme[34]. Successful study should allow, *inter alia*:

- "internationally-mobile students [.....] to transfer from one **IB** school to another;
- students who remain closer to home [to] benefit from a highly-respected, international curriculum;
- [award of] a common [.....] university-entry credential for students moving from one country to another;
- [a] share [in] an academic experience that emphasises critical thinking, intercultural understanding and exposure to a variety of points of view;
- the [promotion of] values and opportunities that will enable [students] to develop sound judgement, make wise choices, and respect others in the global community;
- [the acquisition of] skills and attitudes necessary for success in higher education and employment."[35]

The emphasis on 'internationalism' and 'intercultural understanding', recognising international mobility amongst an **IBO** clientele, centralises the importance of language development through promoting communicative competence in more than one. Thus, a Baccalaureate curriculum requires study of six discrete, 'traditional' or "academic" subject-groupings, including at least two languages at *Higher* or *Standard Levels.* The design encourages cross-curricular interlinking, whereby knowledge and skills acquired in one domain may influence learning in another. Obligations also include the complementary study of *Theory of Knowledge*, the completion of 150 hours of personal *Creativity*, physical *Action* and community *Service*, as well as an individual research project, or *Extended Essay*[36].

The domains researched are:

- *Group 1* or *Language A1*, as literary programmes for "encouraging students to maintain strong ties with their own cultures", normally a home-language, the dominant language of the social environment, or the language in which most teaching and learning take place;
- *Group 2* or *Language A2, B,* or *Ab Initio*, as second languages, ranging from foreign language acquisition by beginners in *Ab Initio* curricula, to the development, or enhancement of bilingualism under *A2* programmes[37].

For the award of *Diplomas*, candidates must offer satisfactory work in at least three and not more than four *Higher Level* subjects, with the remainder at *Standard Level*[38]. Evaluations are numerically graded for each domain, with aggregated scores falling within a 24 to 45 point-range[39]. In this design, no permanent boundaries determine 'passing' or 'failing' grades for any given subject, since all component assessments are subject to formal, *post*-examination moderations, establishing grades anew for each examining session, according to fixed criteria[40]. In most cases, boundaries vary slightly over time, across each group of assessments[41].

For any language and level, assessment material combines work sampled from the span of a course, and point-in-time, external examinations[42]. The inclusion of the former under *Internal Assessment*, produced in varying situations during the final year of instruction, is held to strengthen construct validity, through close linkage to the aims and objectives of the *Diploma Programme*.

The Principles of Moderated Assessment and Evaluation

Moderated assessments and graded evaluations are criterion-referenced for all components of **IBO** languages programmes, in an approach seeking consensus amongst trained assessors and evaluators and eschewing positivistic referencing to fixed, pre-determined norms. The aim is to ensure that similar standards pertain world-wide across time and across the work of all *Internal Assessment Moderator, Assistant Examiner,* and *Team Leaders*[43]. Statistically-derived adjustments in scores ensure consistency. However, ultimate judgements of quality are determined by *Chief Examiners* at *Grade Award Meetings*. Their objectives have been clarified by research through observation. They are to:

- consider teacher and examiner comment for the previous assessment sessions;
- review the procedures and outcomes of these sessions;
- assess statistical information derived by **IBCA** from the relevant examinations, prior to meeting;
- reconsider and evaluate representative samples of candidate work;

- establish grade-boundaries at three points, transformed into numerical evaluations, derived from published, graduated, assessment-criterion descriptions;
- calculate mathematically remaining grade-boundaries to establish incrementally-equal groups of scores;
- apply grade-evaluations thus derived, across an entire population of candidates;
- match samples, totalised per candidate and session to overall evaluation criteria for relevant language categories and levels [44].

For the **IBO**, *Chief Examiner* moderation of sampled assessments by *Assistant Examiners* and *Team Leaders* is a reliability measure, "achiev[ing] the required degree of consistency among assessors of the same subject"[45].

Assessment components are varied, in order "to acknowledge both the content and the process of academic achievement and to take into account different learning styles and cultural patterns"[46]. They include written, end-of-course and point-in-time examination, as well as specialised forms suiting a given subject[47]. Coursework over the period of instruction may be internally-assessed by a candidate's teacher, with **IBO** moderation of sampled productions. For each session, *Chief* and *Deputy Chief Examiners* establish final evaluations at *Grade Award Meetings,* with *Teacher-Observers* in attendance, assessing transparency and regularity of procedure. Evidence available at these meetings includes all examination scripts, teacher comments on the papers set, *Assistant Examiner* reports, notifications of special, irregular or unusual circumstances, samples of statistical data from previous sessions, and the assessment and grading criteria. Awards are open to appeal, whereupon a candidate's productions are fully re-assessed by further examiners, in replication of all procedures[48].

In separate **IBCA** actions, all component evaluations are arithmetically aggregated by weighted value for the relevant scheme. Totals are transformed into final, numerical grades in a straightforward, mathematical exercise. Judgements should accord with *General Grade Descriptors* and score-conversions, tabulated to illustrate grade-boundaries within each programme. Commented summaries of performance are published in subject-specific *Reports*, composed by *Chief Examiners.* These include the descriptors, tables and results for any given session in any given domain[49].

Such bi-annual publications provide evidence for moderating and evaluating processes, revealing **IBO** attitudes and values amongst their authors, as the organisation's representatives[50]. The *General Grade Descriptors* contained within, provide equivalences of *Diploma* grades to aggregated scores, based on a scale rising by integers from a minimum zero to a maximum seven[51]. Also provided are tables for converting individual scores for each discrete assessment into similar seven-point graduations, determined for each component[52].

Within the latter, grades result from matching samples of candidate work to sets of scaled, descriptive criteria, with the "notion that an aggregated score equivalent to Grade 4 represents a passing level in each of the six subjects"[53]. Scores of zero apply when none of the described criteria are satisfied. The procedure serves to guide and check *Moderator* and *Chief Examiner* interpretations. Ultimate overall grades, aggregating individual, componential scores, are adjusted by percentage-value weightings for each, according to programme schemes[54]. However, neither low scores in any single component, nor awards of lower grades in any given domain necessarily jeopardise a full *Diploma* award. 'Compensation' with higher grades in other subjects may suffice to establish satisfactory totals[55].

Language Groupings in the *Diploma Programme*

As stated, **IBO** Language Programmes are classified in two major categories: *Group 1*, or *Language A1*; and *Group 2 Languages*. The latter are further categorised by three discrete, though continually-graduated and implicitly-interlinking levels, as *Languages A2, B* and *Ab Initio.* For diploma purposes, all candidates are assessed in either one *Group 1 Language* and a differing *Group 2 Language*, or alternatively, two differing *Group 1 Languages.*

Excluding *Ab Initio Languages*, there are no published corpora of lexical items or linguistic structures demarcating curriculum content. Thus, there are no overall 'standards', external or internal, by which given languages and levels are categorised as distinct and defined in range, although choice is available in traditionally-classified lists[56].

Hence, 'English' is not differentiated as 'American', or 'British'; 'French' as 'Belgian', 'Canadian', 'Swiss', 'metropolitan', or other[57]. Indeed, with *Languages B*[58], it is recommended that:

> "in the case of languages spoken in more than one country (such as English, Spanish, French, Portuguese for example) candidates should be exposed to a range of varieties wherever possible"[59].

Furthermore, in oral or written production:

> "candidates may use the variety of the language with which they are most familiar. However, they should be consistent in their use of the language."[60]

Such statements contain implicit rationales, warranting comment as evidence of institutional understandings in assessing and evaluating authentic language-productions. Political and socio-linguistic recognition of different forms, sanctioned as 'official', reference-languages for use in internationally-recognised entities or 'nation'-states, may be inferred. Nonetheless, the **IBO** neither defines other standards, nor specifies norms as comparators for positivistic evaluations. As noted, exposure to linguistic variety is explicitly recommended for inclusion in teaching programmes[61]. Relationships between such occurrences and the situation pertaining to French are unclear for the research. They will however be relevant to the analysis of assessment data, described and discussed in Chapter 6.

Diploma Programme courses are broadly differentiated, *Group 1* being a "literature course for native, or near-native-speakers"[62]. In distinction, *Group 2* is defined as predominantly 'language-based', though including the study of literature in varying amounts and for varying purposes, either broadly aesthetic, or purely linguistic, rather than exclusively 'literary'. It is a 'second', (or possibly, a 'third'), and 'foreign' language programme[63]. *A2* and *B Languages* are available for assessment and certification at *Higher* and *Standard Levels*, whereas *Ab Initio* Languages (for foreign-language beginners) are only available at *Standard Level.*

The classifications are not sharply distinguished, discrimination being provided by statements of "needs" for differing students. For example, *A2* candidates are typified as:

- "bilingual students who are capable of studying both of their languages as languages *A1*, but who, for various reasons, prefer not to study two languages *A1*;
- bilingual students who study the better of their languages as language *A1* and require a course of study to bring the other language up to a similar level [with examples of 'typical' contexts given];
- those who have lived for a great part of their lives in a country where the target language is spoken and have gone beyond the foreign learner stage, but are not considered native-speakers of the language;
- those who have been educated throughout the secondary level at a school whose working language is not their native language [examples given]; such students will have surpassed the foreign learner stage, whilst not being considered native-speakers of the language."[64]

For *Ab Initio Languages*, students are typified as:

- "those who have had little or no opportunity for foreign language study in their earlier education and are therefore unable to fulfil **IB** diploma requirements for *Group 2*;

- those who are interested in learning a new, foreign language as a part of their **IB** diploma, possibly in addition to language *A2* or *B*."[65]

Confusions in unambiguous categorisation by student "need" can arise[66]. Revealingly, the **IBO** exhorts teachers and school co-ordinators to display good faith in registrations at particular language levels. Awareness of problems of discrimination, validity and reliability in assessment and evaluation that ensue from inappropriate behaviour is suggested. Hence:

> "teachers and **IB** co-ordinators should ensure that, as far as possible, students are following the course which is most suited to their needs and which will provide them with an appropriate academic challenge."[67]

No further control for this aspect of teaching and learning, ensuring respect of recommendations, is provided. Thus the aims and objectives of discrete teaching, learning and assessment, distinguishing both programmes and levels within programmes, are partially left open to individual interpretation. For **IBO** member-schools, good faith in responsibly administering 'appropriate' curricula and in appropriately grouping students by course or class, may only be presumed.

The *Language B* Programme[68]

Within *Group 2 Languages*, the *Language B* programme is the largest in numbers registered for assessment[69]. It is defined as:

> "a foreign language learning programme designed for study at both higher and subsidiary levels by students with previous experience of learning the language. The main focus of the programme is on language acquisition and development."[70]

Language B explicitly includes the study of "literary" texts amongst others, as "an important part" of the process[71]. Typical students are those who have "already studied the target language for between two and five years immediately prior to the beginning of their **IB** course"[72].

Detailed pedagogical, curricular and assessment perspectives are outlined in the *Guide to Language B*[73]. Given the research themes, salient features may be summarised as follows[74].

Language use is typified by "communicative" production, focussing "principally on interaction between speakers and writers of the target language"[75]. The most significant aim therefore promotes situated use of given languages within contexts defined as "social", "academic" and "cultural" under programme *Objectives*[76]. Curricula and pedagogies should expose students to "a wide range of oral and written texts of different styles and registers", with recourse to "authentic materials [.....] wherever possible"[77], and maximum use of the target language.

These terms are not further explained, though analysis of cross-referenced examples of institutional usage partially clarifies meanings. Indeed, observation of regular **IBO** teacher-training sessions founds some key assumptions. Through use of "authentic materials", teachers are referred to texts produced for an audience and readership of 'native' speakers of the target language. Such documentation is specifically unadapted for pedagogical or assessment purposes[78]. "Maximum" target-language use promotes given languages to the greatest extent feasible, as unique media for classroom instruction and interaction. Monolingual and communicative environments are to be created and sustained through such **IBO**-recommended approaches to pedagogy and learning[79].

Within monolingual channels of communicative interaction, the development of skills in listening, speaking, reading and writing should be as integrated as possible. No hierarchy of importance should justify weightings for evaluating discrete areas of knowledge or skills. Emphases in learning should be "equal", though in *Language B* assessments, 'equality' of value is absent[80]. Likewise, assessment-tasks should integrate as many skills as possible, with speaking both assessable and continually-assessed through participatory, monolingual, classroom activity. As far as feasible, structural features of language acquisition and development should be contained within materials presented for learning and assessment. Aims, objectives, content, assessment and evaluation criteria should be made freely available. For transparency and effectiveness in learning, students should be encouraged actively to participate in all procedure and regularly to assess their own progress[81]. Ultimately, for enhancing motivation and commitment, learners should assist in choosing the texts, topics and activities of their own curriculum[82]. As successive chapters show, key elements of 'authenticity' are prefigured in this distinctive, communicative philosophy of language teaching, learning, assessment and evaluation.

Key aims in **IBO** philosophy relate to components of authentic language use, investigated through the research. For convenience, they may be summarised as promoting:

- accurate and effective communication with others through target-language use in speech and writing;
- transactionally and socially-contextualised communication;
- learning that facilitates use in employment or leisure-time contexts, and effective further study;
- language-use that integrates "insights into the culture of the countries where the language is spoken";
- opportunities for "enjoyment, creativity and intellectual stimulation" as motivating activity[83].

These aims are specified in detail through explicitly and discretely-assessable *Objectives*. For assessments and evaluations, they guide rubric and task-design, and are broadly categorised in three groups as: *Social, Academic* and *Cultural*[84].

Social Objectives are commonly defined at *Higher* and *Standard Levels* as demonstrations of abilities "to respond to the complex demands of day-to-day communication". The recognition of implicit meaning and attitude is isolated for *Higher Level* assessment only. Together however, they relate to the aim of transactionally and socially-contextualised communication with others, with programme requirements specified as:

- "obtaining information from written and oral sources;
- processing and evaluating information from written and oral sources;

- communicating or corresponding with users of the target language in both formal and informal situations;
- making social or professional contacts with people who live and work in the country or countries concerned;
- expressing views and opinions on issues of general interest;
- expressing feelings."[85]

For target-language assessments, *Academic Objectives* relate the concerns of 'self' to those of 'others' by requiring demonstrations at both *Higher* and *Standard Levels*, of abilities to:

- use spoken and written language with accuracy and variety;
- respond with understanding and appropriacy to spoken and written language;
- enter into discussion with the expression of opinion[86].

At *Higher Level*, the repertoire is extended, including concepts and modes such as:

- 'sensitivity' to the spoken and written language;
- appropriate response to 'authentically' academic situations[87];
- debating and the defence of opinion[88].

Here, authentic expression relates to usage in its widest senses, through communicative mastery of a 'working language'.

Finally, in social interactions and varied readings employing target-languages as communication media, *Cultural Objectives* highlight the internationalism of **IBO** programmes. They relate 'others' to 'self' by requiring *Higher* and *Standard Level* assessments of demonstrated "awareness and appreciation of the different perspectives of people from other cultures", together with the "understanding of how language embodies these differences"[89].

The *Syllabus Outline*[90], advises teachers to devise their own curricula by level, based on appropriate objectives and requiring readings and analyses of a wide range of texts "of their own choosing – written and spoken, literary and non-literary". Programmes should explore three, all-embracing themes: *Change*, *Groups* and *Leisure*[91], including integrated, yet systematic presentations, development and revision of appropriate grammatical structures and vocabulary, with "equal emphasis" on "text-handling, written production, listening and oral" skills[92]. Many exemplars are given as pedagogical suggestions, though none are prescribed[93].

The Structure of Assessment in *Language B*[94]

A single design applies to all **IBO** *Languages B*, at both *Higher* and *Standard Levels*. It comprises an examination with two written components, each requiring completion in 90 minutes of supervised, silent and independent, point-in-time effort. The first paper, produced and assessed externally with all tasks and responses in the target language and dictionary use excluded, is *Paper 1: Text-Handling.* No choice is available to candidates. At *Higher Level*, this requires reading four "authentic" texts on course themes, including one "of a literary nature"[95]. At *Standard Level*, the requirement is reduced to three, without necessarily including literary extracts[96].

The second component, comprising similar general rubrics, is *Paper 2: Written Production.* It requires a single composition, selected in response to one of six tasks, with production of a minimum 400 words at *Higher Level*, and 250 words at *Standard Level.* Subjects relate to defined course themes (at *Higher Level*, these include literary themes). Each necessitates a discrete genre to ensure cultural and linguistic appropriacy, in response to defined readerships. One task requires reference to candidate readings in the target language, and another to the theme of one of the texts presented in *Paper 1*[97].

Aural and oral components are integrated within arrangements for *Internal Assessment*[98]. Samples of differing forms are produced in a minimum of four activities at *Higher Level*, and three at *Standard Level.* These should vary, being prompted, facilitated and assessed by teachers during the final year of instruction[99]. At least one must involve group interaction (paired if no more than two candidates are available), one a response to aural stimuli (such as radio, television or cinema broadcasting), and at *Higher Level,* one response must refer to a literary text studied. These requirements may be combined in a single, common exercise. No minimum time duration is specified, and productions may be spontaneous or prepared[100].

Formal oral presentation, with related interview and general discussion, clearly recorded on audio-cassette and dispatched to **IBCA** for moderation, must constitute one activity. For others, examining centres retain written summaries and descriptions of circumstances, together with assessments by the teacher-facilitator concerned. Quantitatively-scored evaluations of these assessments are dispatched to **IBCA** and included within the estimation of predicted, final grades. Whenever a discrepancy emerges from moderation, or results indicate wide variation from teacher-assessor predictions, these reports serve as further evidence for moderating final component grades.

The teacher completes assessment according to published criteria. It requires recording of prepared, uninterrupted, individual presentations by candidates of two to three minutes' duration, on a topic of their choosing, but guided by teacher-*Internal Assessors* for appropriateness within the **IBO** scheme. The related interview ensues, of four to five minutes' duration and followed by an unprepared discussion of a more spontaneous and general nature, up to four minutes in duration. The whole should total ten to twelve minutes of speaking and listening and is dispatched to **IBCA**, after internal assessment, for external moderation.

With later discussion of authentic language use in mind, it is important to note that possibilities for choice are broad, being largely determined by candidates themselves, following their own interest and concerns[101]. The *Internal Assessor's* role is limited to that of guide and facilitator, ensuring that the rubrics for *task*, (rather than the detailed choice of *content*) are respected, in order to allow the fullest application of the relevant assessment criteria. No other rationale for encouraging such candidate 'empowerment' is given[102]. All assessments are completed internally and moderated externally.

Examination Design and Standardisation

IBO documentation researched comprises[103]:

- *General Instructions* for examination production for all languages at all levels, for sessions in May and November 2002;

- *Paper Specific Instructions* for *Language B, Higher* and *Standard Levels, Paper 1: Text-Handling* and *Paper 2: Written Production*, for the same sessions;
- *Checklists* for reporting and evaluating conformity with **IBO** criteria in examination production;
- *Standardiser's Guidelines: Language B*, in most recent draft.

In addition, archived drafts of examinations and correspondence between **IBCA** personnel, examination-designers, standardisers and producers for the May 2001, *French B* examination, provided further data on which the research is based[104].

From *General Instructions* for examination production, it is evident that two imperatives constrain administrative procedure: the requirement that all work be prepared to exacting schedules, ultimately determined by the timings of relevant examination sessions; and the maintenance of confidentiality and security in an international organisation conducting much of its business at a distance, via differing forms of correspondence, including post, telephone, fax and other electronic means[105].

It is clearly stated that in academic content, examinations must "adhere[.....] to the criteria laid down in the relevant published guide", and may also be noted that task-design proposals implementing this formal protocol are not proof-read[106]. They are nonetheless, subject to comment by external, examination *Standardisers* and the **IBCA** *Subject Area Manager*, whose briefs include checking and ensuring conformity with *Subject Guide* statements and the relevant *Assessment Criteria*: key documents to which the entire procedure of examination production is referenced[107].

Standardisation is the responsibility of third-party, external examiners or *Standardisers*, who typically for more commonly-assessed languages are native-speakers and professional, university-level academics[108]. The procedures followed by designers and standardisers are further defined in **IBCA** checklists, assessing comprehensiveness and conformity to organisational policy. Listed in detail are concerns for accurate, legally-permissible reproductions of "authentic" materials[109], as well as for congruence with **IBO** programme requirements and the organisation's chosen formatting styles for publications[110].

Nevertheless, occasional editing of authentically-sourced materials, either for reducing the length of reading texts to suit constrained, examination settings, or for imposing conformity to *Subject Guide* and *Assessment Criteria* requirements, needs inspection[111]. In the editing process, *Standardisers'* duties are explicit:

> "The standardiser will comment on the suitability of the papers and ensure that similar approaches and levels are guaranteed in all languages. [.....] The standardiser will not proof-read the examination papers, nor will he/she be in a position to 'correct' items. His/her comments are suggestions which you may or may not decide to incorporate." [112]

Further relevant issues concern exceptional cases, where it is explained that breaches of internal administrative security may lead to complete rewritings of all items affected. Similarly, all relationships between examination-designers, schools, teachers and potential examination-candidates must be declared to **IBCA**, prior to the assumption of duties as designers.

In *Paper Specific Instructions* for *Language B*[113] and *Checklists* for evaluating examination conformity with criteria[114], examination-designers should also remember that general parameters and purposes for the programme require the following to be respected:

- "the [....] course is designed for students who have studied the language for between two and five years prior to the beginning of their **IB** course"[115];
- "the same level of sophistication cannot be expected from *Standard Level* candidates as can be expected from *Higher Level* ones"[116];
- "the format of the examination papers is the same for both *HL* and *SL* [.....]. However, the choice of questions should reflect the difference of expectations between the levels"[117];
- "the link [between one, two or three of the tasks set in *Paper 2*] with [the themes of reading material in] *Paper 1* should only be tenuous in order not to disturb or frustrate candidates"[118];
- examiners should ensure that tasks are sensitive to the international context of **IBO** programmes and examinations in that they should avoid causing "offence" in "social and political contexts which have different religious and moral beliefs and social conventions"[119];
- "Literary questions must be worded in such a way that any text studied could be used to illustrate the answer. However, questions which are so general that they could be easily rehearsed beforehand must be avoided"[120].

The *Checklist* adds the criteria that each task set:

- "has been narrowed down"[121];
- "is meaningful"[122];
- "can be completed in 1hour, 30 minutes"[123].

In addition, the **IBO** exhorts designers to recognise the organisation's commitment to internationalism through "set[ting] a wide variety of questions which will be accessible to candidates from differing backgrounds", though differences are neither specified, nor given their extreme diversity, are they likely to be specifiable[124]. The intention is "not to limit a candidate's choice of written tasks to only one or two"[125]. As illustrations, suggestions and examples of possible tasks are listed, with general advice that designers respect authenticity as a form of 'naturalness', though avoid tasks requiring responses in "dialogues and conversations", since in written productions, these "can turn into artificial activities"[126]. Hence, all tasks must provide:

- "a context;
- the type of text which is expected (e.g. a letter, an article, a report);
- the audience;
- some indication of the type of register (even though it may be implicit)."[127]

Tasks at *Higher* and *Standard Levels* should be differentiated by "suitability". The concept is exemplified by letters "about holiday plans", deemed appropriate to *Standard,* but not *Higher Level*[128]. Nonetheless, each task should stimulate language-production that permits application of the highest levels of all assessment criteria[129]. For any given criterion, oversimplification and restriction may limit maximum, attainable scores to less than 100%. Notably in this context, a rationale for offering task-choices to candidates is implicit in the requirement that "questions cover a **range** of interests, [avoiding] gender bias", and that they be "relevant and interesting to a 17 – 18 year old student."[130]

From these *Paper Specific Instructions*, it is evident that examination tasks are intended to stimulate communicative, authentic language-production. Candidates' chosen responses should be linguistically and culturally contextualised in ways appropriate to, and thus determined by specific task-designs. Implicitly, through covering "a range of interests", and being "relevant and interesting [for] a 17 -18 year old student"[131], designers should encourage motivated response, requiring expression through writing in a given target language.

Indeed, the spirit of a 'philosophy' of authenticity is detectable in explicit pleas for 'realism' in task-setting (albeit for *Paper 1*), where instructions state that:

> "examinations inevitably involve a degree of artificiality, but the necessary conventions of examination tasks should mimic operations carried out in real life as far as possible.
>
> As far as possible, all tasks set in the text-handling paper should be **realistic**. In other words, they should be operations that the average educated reader should want or need to perform in order to understand the passage properly, or to make use of the passage successfully."[132]

In ‘narrowing down’ and creating ‘meaningful’, or ‘realistic’ tasks, designers are granted discretionary powers, significantly ‘positioning’ themselves and candidates, and influencing the authenticity of communicative language use in response[133].

However, possible imbalance, introduced through the particularities of perspective, understanding and choice amongst individual examination-designers are recognised in statements of procedure. Hence, following comment and recommendation by *Standardisers* and the *Subject Area Manager* in accordance with their briefs, **IBCA** permits further remodellings for finalising task-design.

Here, it should be noted that administrative advice for working to **IBO** *Guidelines*, does not reproduce the data of *Subject Guides*[134]. Instead, instructions to *External Advisors* and *Standardisers* elaborate and contextualise organisational understandings within frameworks offered by the *Guides*. To prefigure the discussion of Chapter 6 and from comparing primary evidence, it may be seen that documentation for internal use refines understandings, adding greater detail regarding specific, institutional circumstance and modifying often implicitly-understood terminology, as employed by the organisation’s personnel. Thus the following are significant:

- differentiation between *Standard* and *Higher Level* is one of “sophistication”, rather than anything else[135];
- there is an explicitly-stated “difference of expectations between the levels”, although both differences and expectations remain implicit[136];

- the requirement that examination-designers and standardisers consider task ‘suitability’ according to specified levels, underscores implicit differences in expectations[137].

The *Guidelines* issued to *Standardisers,* together with the associated *Checklist*, add further criteria whose satisfaction is required in preparing examinations. In part, these resolve certain ambiguities present in the *Paper Specific Instructions* for *External Advisors*, or examination-designers. In the general introduction for instance, *Standardisers* duties of all *Language B* examinations, should ensure:

- “conform[ity] to the same rules and regulations”[138];
- “a comparable level of challenge to candidates irrespective of which *Language B* they study”[139];
- conformity with the *Paper Specific Instructions* under which examination-designers perform their duties, yet with the distinction that *Higher Level* papers “should demand a higher level of linguistic ability and sophistication [than *Standard Level* papers]”[140].

The guidelines for *Paper 1* need not concern us for reasons previously outlined[141]. Those for *Paper 2* reiterate concerns for equity in offering opportunities to candidates to meet all assessment demands, regardless of language studied.

However, it should also be noted that the most recent draft for *Standardiser’s Guidelines: Language B* introduces an additional notion, not present in the *Paper Specific Instructions* to examination-designers. This requires *Higher Level* papers to “demand a higher level of

linguistic ability[142] and sophistication" [than *Standard Level* papers]. Appropriately distinctive "abilities" in language are left undefined[143].

In more detailed listing of requirements to be checked through standardisation, the notion of linguistic differentiation between *Higher* and *Standard Levels*, should clearly allow examinations to be, *inter alia*:

- "accessible even to weaker students while allowing stronger students the opportunity to excel;
- appropriate to the level (*H[igher] L[evel]* questions should overall be more challenging than *S[tandard] L[evel]* ones)"[144].

From inspecting data for the design, standardisation and production of May 2001 papers for *French Language B*, and in interviewing the **IBCA** *Director of Assessment* and the *Examination Papers Officer*, the following were relevant[145]:

- in the case of English and French, examination-designers (or *External Advisors* on examination design) and *Standardisers* are always native-speakers of the respective language;
- given the 'correct' application of the design, standardisation and production criteria, longitudinal standardisation of similar examinations over time is not significantly meaningful for **IBCA** validation purposes;
- the entire process on this occasion had taken fifteen months, with draft examinations for *Paper 2* amended twice by two different **IBCA** officers, prior to proof-reading for publication;
- amendments and revisions concerned questions of grammar, vocabulary usage, contextualisation within a cultural specificity,

the 'realism' or perceived 'artificiality' of tasks proposed, and the ensuring of 'appropriate' differentiation between the demands and expectations of *Higher* and *Standard Levels.*

For deeper research into authenticity, it is the latter item that requires further investigation, with examples drawn from empirical data. For the design, standardisation and production of the May 2001 examinations, specific instances have been isolated and described. They are presented, analysed and discussed in Chapter 6.

Assessment and Examination Administration

In the *Guide to Language B*[146], there is common definition of assessment procedures, with structural and analytical categorisations, classifying sets of descriptors as graduated assessment criteria for *Higher* and *Standard Levels.* For oral production, they are in overview:

- *"Criterion A: Task/Message:* The effectiveness of the speaker in completing the task when communicating the required message;
- *Criterion B: Interaction:* The effectiveness of the speaker in maintaining the flow of the discussion;
- *Criterion C: Language:* The accuracy, appropriateness and fluency of the language used."[147]

No rationale for this tripartite categorisation is made explicit[148], although the three presented are further 'illustrated' with exemplary questioning. This typifies the assessment of *Task/Message* by assessors and interlocutors, as consideration of:

- *Overall performance* (as the "interest" value of the content of productions, presumably for interlocutors and assessors);
- *Task* (as the degree to which tasks have been completed);
- *Message* (as clarity of message and appropriateness of response);
- *Ideas* (as illustration of the ideas and arguments presented, as well as of relevance, interest and convincingness, assumedly again, for interlocutors and assessors)[149].

The following sub-categorisations should be assessed for *Interaction*:

- *Overall performance* (as the "liveliness" of communications);
- *Interaction* (viewed as degrees of contribution in exchanges);
- *Coherence* and *fluency* (neither further defined, nor exemplified);
- *Responsiveness* (as degrees of comprehension of spoken language and "appropriateness" of response)[150].

The assessment of *Language* requires consideration of the following:

- *Overall impression* (focussing on the fluency of communications);
- *Vocabulary* and *register* (as degrees of appropriateness and variety in usage);
- *Accuracy* (focussing on the variety and accuracy of grammatical structurings);
- *Pronunciation* and *intonation* (as the contribution to fluency provided by these categories)[151].

In written language-production, a similar, tripartite specification is also employed, categorising assessment criteria as *Task and Message; Presentation;* and *Language.* Again, these are defined in common at *Higher* and *Standard Levels*[152]. In overview, they are:

- "*Criterion A: Task/Message:* The effectiveness of the writer in completing the task when communicating the required message;
- *Criterion B: Presentation:* The organisation and cohesion of the text;
- *Criterion C: Language:* The accuracy, appropriateness and fluency of the language used."[153]

No explicit rationale for this further, tripartite categorisation is provided[154], although the three dimensions are again, better understood through reference to exemplary questions categorising assessment in *Task/Message* as:

- *Overall performance* (or subjective assessments of success in task-completion and of clarity and effectiveness in producing a message);
- *Content* (as degrees of comprehensiveness in presenting information and the clarity of the ideas chosen);
- *Task relevance* (neither further defined, nor exemplified);
- *Ability to convince* (as assessments of the presentation of arguments and responses to the expectations of given readers, presumably including assessors)[155].

The assessment of *Presentation* requires readers to consider:

- *Overall presentation* (as clarity and effectiveness in presenting ideas and/or information);
- *Paragraphing* (as the degree of contribution by this feature to the development of ideas presented);
- *Cohesion* (as the linguistic variety and appropriacy of the cohesive devices employed for maintaining continuity in written expression);
- *Register* and *style* (as degrees of appropriacy to topics and tasks chosen)[156].

The assessment of *Language* requires consideration of:

- *Overall impression* (focussing on the fluency of language employed);
- *Vocabulary* (as degrees of appropriateness in range, and of accuracy in usage);
- *Grammatical accuracy* (focussing on the variety and accuracy of grammatical structurings);
- *Intelligibility* (as the contribution to clarity and accuracy of manuscript writing, focussing on orthography)[157].

In assessing speaking and writing with **IBO** criteria, a recommended approach is first to consider *Language*, then *Task and Message*, and finally *Interaction* (for oral performance), or *Presentation* (for written production)[158]. In all cases, productions should be compared with the general descriptors for the lowest levels of performance, and if inappropriate, those for the next level, and incrementally, up to the most appropriate description possible. On selecting general descriptions that 'fit' best, judgements should be confirmed through considering the detailed, further descriptions, modifying eventual scores accordingly for each category[159]. An element of interpretative subjectivity is allowed in the process. In finalising scores within each, generally-described, two-point category range, the **IBO**'s informal suggestion is that *Moderators* and *Examiners* may compensate either a tendency to severity or generosity in one criterion, with its opposite in another[160]. Ultimately, a total score of a maximum thirty of points is awarded through aggregating the three, discrete criterion-scores, (each with a maximum of ten), and recorded on **IBCA** *pro-formae* designed for the purpose[161].

Certain copies, if problematic in examination, are thus processed through a number of listenings and readings, with unresolved cases dispatched as "problem cases" to *Moderators,* or *Examiners* responsible for co-ordinating examiner teams, for further assessment. The criteria for such cases, with examples of 'commonly' encountered problems and procedures for their resolution, are published by the **IBO** (1996a), in the appropriate administrative documents[162].

With assessments completed for each session, *Subject Reports* are drawn up by *External Moderators* and *Assistant Examiners*, together with *Reports* on the performance of all candidates from particular centres requesting these. *Subject Reports* record:

- the parts of the programme that candidates appeared to find difficult;
- the levels of knowledge, aptitude and comprehension displayed;
- candidate strengths and weaknesses in responding to individual tasks;
- the type of help and advice that teachers should give future candidates;

- any confidential description of problems and comments relating to schools that appear to be in serious difficulty in delivering the programme;
- points for discussion at *Grade Award Meetings*[163].

For the latter, observations are classified as:

- Remarks on candidates' examination techniques, with suggestions for improvements;
- Remarks on the presentational quality of candidate work, with suggestions for improvements;
- Totals for the numbers of candidates attempting each task;
- Remarks, where applicable, on the overall performance of candidates per section of the examination, including preferential choices;
- Analysis and evaluation of candidate performance according to the assessment criteria and tasks attempted;
- Recommendations and advice for future candidates[164].

All reports are dispatched to *Chief Examiners* for further consideration through moderation, and for summary by *Chief Examiners* contained in the *Subject Reports.*

Weighted Values for Listening and Speaking, Reading and Writing

Papers 1 and *2,* for reading and writing, are weighted at 40% and 30% of the total score, respectively. Besides these *External Assessment* components, a remaining 30% is devoted to the *Internal Assessment* of listening and speaking[165]. This distribution is held as acceptably reliable and significant for international, tertiary-education establishments accepting **IBO** programmes as 'high-stakes' and credible for entrance purposes. Indeed, *Chief Examiners* are employed as overall supervisors by **IBCA**, partly to establish and maintain credibility for the academic standards of programmes and assessments, as adequate for uncontested university recognition. Additionally, the arrangements allow both greater reliability in grading across differing administrations of assessment sessions, and standardisations of task and examination design that do not require further moderations for ensuring compatibility with requirements[166]. The entire structure illustrates the influence of governments and universities as 'clients' of **IBO** evaluations, expressed not least through the composition of the organisation's governing bodies and choice of *Chief Examiners* and *Standardisers*[167]*.*

In researching authentic language use, the distribution of percentage scores allocated to each major assessment component, serves to indicate **IBO** perspectives on relative quantifications of 'value' for discrete language skills and knowledge, even though these are intimately interlinked and effectively inseparable. Emphases are significantly discernable.

Internal Assessment, valued at a maximum of 30% covers listening and speaking, though without equal weightings of 15% for each componential skill. The three sets of assessment descriptors, categorised as *Task/Message*, *Interaction* and *Language*, implicitly place greater value on active, language-production, rather than its reception. Thus when evaluated through matching to given descriptions and converted into scores, speaking ability is promoted at the expense of listening comprehension. As noted, each criterion is discretely and equally valued at a maximum of 10 points, though all are inevitably integrated by situated language use, with value implicitly transferred, and transferable, across criteria. Evidently, *Language* must first be comprehended, if only minimally, to allow appropriate initiations of communicative interchange. Subsequently, in order to facilitate *Interaction* and the communication of *Messages* responding to given *Tasks*, it must be sufficiently comprehensible. The value of comprehension is implicit in descriptors of minimal performance, scored at zero. These relate to samples of language-production in which the level required for description by any higher descriptor has not been attained[168]. Hidden graduations of quality and sophistication are revealed by the description of minimal performance at both *Higher* and *Standard Levels*, in identical terms. Evidently these negative descriptions, detailing the absence of higher qualities, cannot relate to identical levels of attainment. (Indeed, it may reasonably be inferred that minimal performance at *Higher Level*, may potentially provide adequate evidence for scorable performance at *Standard Level*, meriting recognition through awards of scores greater than zero.)

The results of the phenomenon, emphasised by recommended timings for discrete sections of oral assessments, (the first being initiated by candidates with no recourse to listening at all), implies that by value, speaking competence receives greater recognition than listening. Indeed, it could be claimed that listening competence extending beyond the initiation of communication, is only explicitly assessed under criteria for *Interaction*. Arithmetically, listening may earn recognition valued at less than one third of the points available. Regardless of anything other than the most elementary listening skill, language-production represents much of the remaining two-thirds, or more.

Similar imbalances in assessment values for reading and writing may be inferred from the written examinations, evaluated at 70%. The programme defines weightings of 40% for reading, and 30% for writing. However, successful performance in *Text-Handling* evidently requires skill in writing that records both comprehension and appreciation in assessable fashion. Likewise, successful performance in *Written Production* requires reading competence in fully comprehending the implications of tasks set and chosen for response. The true weightings of skills are again relative and ambiguous, with ill-defined privilege granted to well-developed writing in response, and accepted as reliably demonstrating an inferred comprehension. Hence, any irrelevant task-based production may still score greater than zero for its communicative qualities as a message, even if low. Appropriate comprehension of the tasks set may have in no way been indicated, with possibilities created for the inauthentic, non-interactive, prior practising and memorisation of responses, deemed model by candidates or teachers[169]. However, failure to communicate successfully through writing does not evidently expose failure to comprehend either reading texts in themselves, or tasks set for response to reading. Rather it may be attributable to other inchoate, and for assessment purposes, unknowable causes.

As in *Internal Assessment*, each criterion for assessing writing is discretely valued at a maximum of 10 points, though all are inevitably integrated in any situated production of language, with value again, implicitly transferred and transferable, across all criteria. Evidently, *Language* must be of a sufficient level of comprehensibility to allow the communication of *Messages* in response to given *Tasks.* Skills in *Presentation* may be independent of competence in the target language, especially in the appropriate organisation of content within genres that may be common to more than one linguistic culture.

Such imbalance in the weighting of values is once more implicit in the repetition of identical descriptions of performance worth zero[170]. Again, minimal scores at either *Higher* or *Standard Level* cannot necessarily and justifiably refer to identical levels of individual performance. And yet the results of the phenomenon, emphasised by the effects of language quality, are spread across all three criteria[171]. Competence in producing written language thus receives greater recognition than any discrete qualities in task-response, message-construction or presentation. Hence language may be valued, as previously, at greater than the third of points available and explicitly devoted to this criterion.

Consequently, **IBO** assessment criteria positively weight the values of successful production of written language, most notably to the detriment of measurements of quality in listening comprehension. The explicit weightings attached to evaluating certain discrete aspects of language-production, remain ambiguous in rationale, and therefore problematic.

Internal Assessors, Examiners, Moderators and Evaluators

Internal Assessors are normally candidates' own teachers, and *Internal Assessment Moderators*, **IBO** examiners employed to enhance validity through checking conformity of productions to rubrics, thus improving the reliability of results. Examining centres may exercise discretion over the choice and allocation of *Internal Assessment* interlocutors in the best interests of their candidates. In this regard, the **IBO** instructs *Internal Assessment Moderators* to remain as close to original, teacher and *Internal Assessor* assessments as possible, confirming judgements and scorings unless it is wholly evident that these are irregular or invalid[172]. The candidate's own teacher is 'positioned' as normally "the best placed" in the assessment process, through wider knowledge of the personalities and contexts of assessed performances[173]. *Internal Assessment Moderators* should where possible, refrain from recommending changes, focussing instead on establishing reliability measures through comparing centres and teacher-assessors whose candidates' productions they moderate, across the range allocated. Exceptional, or problematic cases are excluded from sampling for further assessment by *Chief Examiners* and are evaluated through separate, supplementary moderations[174].

In assessment, general **IBO** policy is described as a "partnership" between classroom teachers and examiners[175]. Practising teachers are encouraged via postings on the organisation's *Internet* site, and through advertisement in its regular publication, *IB World*, to participate in processes as part-time employees of the organisation. Besides checking academic credentials and school employment affiliations to ensure appropriacy of placement, the organisation maintains internationalism by ensuring variety in the spread of nationalities across its teams of examiners[176]. *Internal Assessment Moderators* and *Assistant Examiners* are recruited under these conditions, with contracts renewable on invitation after satisfactory performance at each examining session. This is monitored and the results sent to the employee concerned. Consistency in applying assessment criteria, respect for organisational deadlines, and satisfactory completion of required reports on the examination component and session in general, as well as of *Individual Subject Reports* at the request of examining centres, are recorded as key elements. All work completed by *Internal Assessment Moderators* and *Assistant Examiners* is supervised by experienced *Team Leaders,* in turn supervised by *Chief Moderators*, *Deputy* and *Chief Examiners* who are required to maintain regular contact with supervisors over the relevant examining session.

For *Group 2 Languages*, *Deputy* and *Chief Examiners* may fulfill all roles simultaneously[177].

Following preliminary, *Team Leader* assessments of sampled candidate work for the relevant examinations, team members receive assessment guidelines from supervisors, reporting from **IBCA** meetings held to determine policy for each session. From an average allocation of up to 50 recordings for *Internal Assessment*, and of up to 150 scripts of *Written Production*, *Moderators* and *Assistant Examiners* are required to send samples of their assessments for further moderation by an examining *Team Leader*, within three weeks of the examination. These samples should represent as broad a range of marks from the allocation as possible, totalling eight recordings for *Internal Assessment*, and a maximum of twenty scripts of *Written Production*. In addition, particular cases of difficulty are reported to *Team Leaders* for discussion and advice, with unresolved problems referred for supplementary assessment and adjudication by the leader concerned. Further assessment and final evaluation by grade on the **IBO**'s seven-point scale, takes place at formal moderation sessions in *Grade Award Meetings*, as described.

Whilst serving teachers are encouraged to fill most posts, it should be noted that the policy of the organisation for recruiting *Chief Examiners* is international, and employment is offered to recognised academics of university, or equivalent standing, whose first language is the language of the assessment and evaluation domain concerned[178].

The Researcher as Employee of the IBO

To conclude description of the context of the research, it should be noted that the researcher has been employed by the **IBO** as *Assistant Examiner for French Language B, Paper 2, Standard Level*, since the inception of the programme in 1996, to date. In addition, from the May 2001 examining session to date, responsibilities include those as *External Moderator for the Internal Assessment Component for French Language B, Standard Level.* From 1997 to date, complementary employment has been undertaken as *Workshop Leader* in regular, international training sessions for teachers either new to, or engaging in teaching the programme for *French Language B, Higher* and *Standard Levels*. Additional roles have covered that of *Teacher-Observer* for the *Grade Award Meetings* for *French Language B, Higher* and *Standard Levels* in December 2000, and for *German Language B, Higher* and *Standard Levels* in June 2001. From inception in September 1999 to completion in September 2002, the researcher has also been a member of the **IBO** *Review Committee*, reviewing the *Language B* component of the *Diploma Programme: Group 2 Languages* in its entirety.

This employment situation has facilitated access to the appropriate sources of data on which the research is based.

PART II

THEORETICAL BACKGROUND

CHAPTER THREE

SIGNIFICANT THEORY: A LITERATURE REVIEW

Preface: Authenticity as Theory and in Practice

Literature investigations relevant to the research are reported in two chapters, over three sections. This chapter presents authenticity as theory, *per se*, together with concepts of communicative and authentic language use in educational contexts. The second concerns authentic assessment in approaching the specification of rubrics, criteria and procedures for measuring and evaluating linguistic performance[179].

Preparing a Review of the Literature

In philosophies of mind and identity, the literature conceptualising authenticity is vast and long in history. Established perspectives in this tradition are reviewed, comparing 'espoused theory' with 'theory in practice', and implying paradigmatic structurings[180]. Indeed, attempts to categorise and describe discrete features of authentic language use, derived eclectically from theory and practice, form one unbroken thread marking out research explorations. As in any linguistic production, second, or 'foreign' language use is assumed partially to reveal the workings of agentive, though linguistically and culturally-socialised 'minds'. They are those of *Diploma Programme* candidates, assessed at particular times[181].

Jean-Paul Sartre's (1946a; 1946b) exemplary, yet sophisticated conceptualisation of authenticity offers a starting point[182]. It is integrative, interlinking ontology and epistemology, the latter emerging from the former. With ethical and axiological perspectives included in the whole, existentialist phenomenology is particularly appealing as a paradigm for referencing key issues in teaching, learning and assessment[183].

However, Sartre's work lacks an extended, systemic account of the significance of language *per se*, and of communication through language as a key mode of authentic, social and educational relations. It offers little analysis of active and interactive language use for structuring human identities. Nor does it describe the components of their construction. Sartrean perspectives provide neither linkage to the traditions of structural linguistics, psycholinguistics or sociolinguistics, nor discussion of these disciplines as approaches to explaining and measuring language-performance. In phenomenological ontology and epistemology, the view that awareness and knowledge inhere in language systems, be they 'native', 'mother', 'first', 'second', 'foreign', 'ancient', 'classical' or 'modern', is not foundational[184].

Nevertheless, existential phenomenology conceptualises authenticity through issues of identity, integrating active intention, either self-oriented or communicative, with propositions of moral value.

Despite the complex unity of Sartrean thought, descriptions of authenticity may be developed from this base. Categorisations of authentic language use may be related to pedagogy, learning, assessment and evaluation. A framework for determining criteria, identifying and facilitating coherent evaluation of language quality through analysis of communicative usage, may be derived from such conceptualisation[185].

Traditions in the Philosophy of Authenticity

'Existential' conceptualisations of authenticity are initially ontological. For Sartre (1946a, 1946b), representing 'being' requires recognising foundational primacy for the existence of a transcendent ego, or individual 'self'. This 'self' is defined as consciousness, aware of its own existence through effects of gratuitous intuition, rather than *a posteriori*, metacognitive reflection. In Heidegger's (1927,1962) terms, it simply 'is there'[186]. Subjectivity as Cartesian *cogito*, exists *prior* to essentialising conceptualisations that are *subsequently* formed as perceptions of location 'there, in the world', and organised *a posteriori*, as 'explanations' for its origins.

This notion of consciousness is anti-deterministic inasmuch as the gratuity of 'selfhood' permits theoretically limitless, 'self'-directed extensions of awareness and understanding. However, the very recognition of 'self' as initially existent simultaneously separates and distinguishes individual consciousness from objects existent in environs to which it is 'external'[187]. Sartre denotes this transcendent, phenomenological '*ego*' as given and 'free', insofar as its existence has neither definable, determining 'cause', nor *a priori* purpose[188]. In existential phenomenology, idealised subjectivism is the ontological foundation for defining and acquiring all knowledge through extensions of 'self'. This applies in educational settings as much as anywhere else.

From idealised subjectivism, Sartre defines authenticity as a dimension of relations, necessarily linking the inner consciousness of existence as 'self', with the apprehendable, phenomenological features of objects, materials, or quite simply 'other', as 'non-self', in a world 'outside'. Socio-culturally and temporally contextualised, 'otherness' is constituted by 'self' through the operations of agentive minds and from sets of sentient perceptions, intentionally-chosen and organised as mental representations. Nonetheless, 'other' remains an inalienable part of a world beyond any gratuitously-perceived bounds of 'self', existent inasmuch as it is perceivable and connected to 'self' through dialectical relations of communication.

Sartre opposes materialist ontologies that reify primary qualities of existence through positivistic measurement by 'norm', or reference to immutable 'laws' discovered as 'science', in order to identify and explain rationally, psychologically, sociologically, or purely linguistically, material origins for human subjectivity[189].

To summarise, authenticity in the primary, Sartrean definition of the term requires recognition of the gratuitous, purposeless primacy of the phenomenological 'self' and of its subsequent limitless growth through intrinsically-extendable activity inherent in all mental processes[190]. Given linguistic and intersubjective collaborations for constructing meanings though communication, individual 'selves' make contact and negotiate with 'other' subjectivities. This phenomenon is especially pertinent in education, wherein an individual's freedom of 'choice' in managing these processes is equally foundational.

For Sartre, all choice is fundamentally 'free'. Simply put, the 'self's' choice of focus in attention is unconstrainable, a phenomenon establishing its 'freedom'. Within temporal flows of existence, unending selections are individually made from ever-present, myriad possibilities[191]. The construction of identity is a perceptual becoming. Existential choice, comprising the conscious directing of future attention, is fundamentally subjective. (In states of non-reflexive awareness, it implies rejecting alternatives through ignoring their possibility as options). No externality serves as prior cause capable of explaining the phenomenon. No effect may annul capacities and possibilities for choice in further, precise focussings of attention. For Sartre, the individual is 'free', though 'condemned' to construct an ever-evolving identity as 'self' through unceasing, effectively unlimited and unavoidable operations of choice. Not to choose is seen as a choice in itself.

In educational contexts, such perspectives on authenticity and on authentic language use, place learners as individual subjects at the centre of all pedagogical relationships (with teachers positioned as co-subjects and co-learners constructing shared, interactive, dialectical meanings). This learning may be assessed, when values are placed upon judgements of outcomes, to be accepted, or rejected by learner-subjects in further acts of choice[192].

The position has received critical attention, notably from materialists, and perhaps most from classical and neo-Marxists, structuralists, cultural and critical theorists. In considering philosophical, rather than communicative, cultural or linguistic dimensions of such critique, Theodor Adorno, presents typical objections[193].

For Adorno (1969), existential phenomenology is untenably idealistic, in that ontology is initially separated from epistemology in *a priori* fashion, despite the contrary insistence of Sartre[194]. Sartrean perspectives are dualistic, distinguishing mind as 'self' from objective worlds contextualising it as ‘other’. All epistemology has been unified *a posteriori,* as a single, subjective and agentive construction developed from pure ontology[195].

Adorno identifies a ‘mistake’ in according ontological primacy to individuals with ‘free’ choice who initiate a communicative dialectic in relationships with ‘other’, thereby enabling subsequent constructions of knowledge. Subjectivism conceptualises identity as transcendent, ‘self’-constructed, and thus ‘independent’ of environments within which individual subjects are located. Adorno reverses Sartre’s position by according ontological priority to the existence of an objective, social, culturally-specific world. This existence is assumed as rationally self-evident, and foundationally given. Accordingly, the psychological worlds of Cartesianism, of Husserlian phenomenology, of Heidegger’s and Sartre’s existentialism lose meaning as private, ideal, independent (or for Descartes, God-‘given’) transcendencies, realised through environmental engagements initiating the shaping of ‘self’. For Adorno, ‘self’ decoupled from its environment, is literally meaningless.

In this Frankfurt School of thought, communication between individuals is definably created and determined by social, cultural and linguistic relations existing prior to any expression of ‘selfhood’. These may be apprehended, typified and subsequently analysed within any particular context and at any point in time. They are exterior and superior to ‘self’, and deny foundational status for conceptualisations appealing to notions of individual transcendence. They anchor understanding in a material world whose mysteries are available for demystification through the progressive and accumulative processes of rational investigation and empirical analysis. All will come to be fully ‘understood’ in an eventual closure of thought, privileging the purposes and methods of positivistic science. For the polemical Adorno, concerns about 'authenticity' are no more than existentialist 'jargon'.

Such critique however, fails to account for the second foundational step of existentialist phenomenology. This builds on recognising the transcendent, Cartesian subjectivism of consciousness, by identifying existent 'selves' as intentional. For growth, they seek dialogic contextualisation in intuitively-perceivable 'worlds' that are representable to 'self'. These worlds are 'exteriority', created through the oppositional dialectics of 'self' and 'other', inherent in all communication. Awareness of 'self' as existent subject, dialectically creates within itself relationships permitting apprehension of the existence of 'objects', and of other 'subjects', in turn perceivable as 'objects'[196]. Adorno ignores the distinction through writing it off as 'word-play'. The dynamics of communicative dialectic, (defined by Sartre as qualities of the simultaneous, dual consciousness of 'self' and 'non-self'), further refines Sartrean ontology, linking it to a distinctive epistemology of identity that modifies his conceptualisation of authenticity. It highlights questions of linguistic production, together with assessments and evaluations of language use, illustrating a key rationale for underpinning the research in this way.

In existential epistemology, it is only through communicative and dialectical relationships between 'self' and 'other', that the former attains meaningful knowledge (rather than intuitive awareness) of itself[197]. Simply put, the contours of 'self'-knowledge are 'constructed' in relation to 'other', and *vice versa* (through initial awareness that knowledge is founded by 'self'). In social contexts perceived through 'self'-chosen focussings of attention, individual subjects entering into relations of communicative dialectic simultaneously serve as 'objects' for any 'other'[198]. 'Self' is understood as initiating relations according to the structuring of its own perceptions, through continual, and freely-chosen engagement with 'other'. To abstain from the 'free choice' of engagement, were this ultimately possible, would be to lose perspective on perceptions of individual existence, and in this sense to cease to exist. For existential phenomenology, clear relationships between teachers and students, between students, assessors and evaluators, are prefigured within this setting. Original choice allows entry into all social relations, maintained and further developed through continuous focussings of attention and engagement[199]. As such they are always, literally 'educational'.

In short, Sartrean authenticity involves recognising ontological, epistemological, ethical and axiological precepts, set within a unified phenomenology. It entails the constructive evolution of identity for 'self' in interactive, dialectical and communicative relations with 'others', through freely-chosen acts socially engaging this 'self', (and also in metacognitive reflection on 'self' as 'other', or as the object of its own attention). When chosen in 'good faith', these acts recognise and respect the similar and equal status of 'other' as alternate subjects in their own right, even if temporally-situated within definable, intersubjective relations of power, embedded within distinctive cultures and societies, and communicable through use of determinate, language systems.

Authenticity in the World of Education

Understanding authenticity as an aspect of identity expressed partly through language use in communication with others is of educational importance. Whether considered in psycholinguistic, sociolinguistic, or purely linguistic terms, significant issues concern the intersubjective and interactive nature of all communicative language reception and production[200]. By definition, language use will be socioculturally-embedded, even within small worlds of those 'second' language classrooms where it is established as the medium of pedagogy and learning[201].

The relationships researched involve microgroups of two, three, four (and in sampled, moderation cases, somewhat more) 'anonymous', individual candidates, communicating with an *Internal Assessor, Moderator,* or *Examiner.* They are framed by socio-culturally accepted, linguistic norms for the given languages through which they are expressed[202]. Such 'standards' permit commonalities for initiating communication between pairs of individuals who may not otherwise know how to understand each other. They also favour non-idiosyncratic assessment of ensuing language-productions[203].

In this context, 'cultures' and their 'norms', or 'standards' may be conceived as mediations of individual minds, as suggested by Lantolf (2000). They shape the mental processing of symbols received as input from socio-temporal and physical environments (albeit through subjectively-determined choices to attend to the initial reception of such symbols). Mental representations are internally organised in subjective ways, and understood as satisfyingly adequate and coherent (thus requiring neither change, nor modification, nor reorganisation). For cultural theorists, these symbols are communicable and simultaneously modifiable through the active agency of individual endeavour. Whether through using tools or thought expressed in language, symbolic intercommunication is fluent. It evolves from interventions indicating personal or mutually-accepted usage and understanding. Despite the particularities of such communication, cultural and linguistic symbols form coherent units of analysis, dialectically linking unique psychological worlds to histories, geographies and societies, as settings external to individuals, yet within which they are located. In totalised sets, representations symbolise apprehendable features of contexts, defining given 'cultures'[204].

Heuristics provide central concepts for socio-cultural perspectives such as Lantolf's, both in communicative praxis contextualised by particular social relationships, and as dimensions of socio-linguistic philosophy. This holds true regardless of the status of any 'standard' language, whether categorised as 'native', 'second', 'foreign', or 'modern'[205]. In particular, language use in such relationships is linguistically-contextualised (even if imperfectly so), through implicit recognition and sharing of canons, be they ideal or 'real', defined and published by bodies such as national academies established for the purpose. Activity takes place and is regulated within situations mutually recognised as viable. Reference to uncontested 'standards' benchmarks the assessment and evaluation of productions and the comprehension these signal. Given the pervasiveness of any language 'norm' in shaping prescriptive, published curricula, whether explicit or implicit, the opportunities for choice and constraint in the range of choices presented, become the focal points of interest.

The Concerns for Pedagogy and Learning

This discussion illustrates why Sartrean approaches were chosen to guide the research. Existential phenomenology is appropriate and capable of illuminating the significance of authenticity as central to communicability in social relationships. Its perspective offers coherent views of sociocultural and sociolinguistic worlds in education, pedagogy and learning of communicative language, as well as in assessing and evaluating its use[206].

For theoreticians of human cognition and learning, communicable 'self'-identification and constructed authenticity in identity lie at the heart of language use. The thought of Piaget, Piagetian, and neo-Piagetian educational psychologists is typical[207]. In Sartrean fashion according to Wood (1988), this claims primacy for an ego-centred awareness, characterised as inherent in the notion of 'self'. Ego-centredness is expressed through 'self-directedness' in which contextualised external reference and behaviouristic 'reinforcement' for accepting its own existence are unnecessary as pre-conditions for learning development. Rather, intrinsically-motivated individuals seek not just to understand 'other', but to extend inner 'selves' through engagement in social and communicative intercourse with such 'other'[208].

Piagetian views on mentation postulate learning as the progressive development of abilities to 'decentrate', or accommodate perspectives alternating with those provided egocentrically, and ultimately arriving at a capability for 'metacognitive reflection', whereby 'self' turns inwards on its own thought-processes as decentrated subject, observing the composition and processes of its own, internally-perceived motion in identity and time, as if from a position of 'other'. Developmental extensions of 'self' permit increasingly-sophisticated engagements with the 'outer' worlds of nature and society[209]. For Piagetians, as for Sartre, motivated, internal, mental activity, directing attention and giving rise to personal experience, is the precursor to all learning.

Sartrean and Piagetian paradigms locate educational development within a socioculturally-contextualised process of enhancing ability in communication, depth and range of understanding through intersubjective, dialogic language, creating spaces for possible 'instruction'. The conceptualisation is further developed in detail by the psychologist and theorist, Lev Vygotsky[210].

Vygotskian notions of 'zones of proximal development', base learning on an individual's prior acquisition and understandings of knowledge, deemed privately and culturally relevant, and motivated for interest. In addition, coherent choices focalising learner attention follow communicative and dialectical interaction with a world perceivable beyond 'self'. Engagement in this world produces understanding with identifiable shape and distinguishing culture, be it linguistic or otherwise[211]. Its features are initially recognised by subjects, and partially, though progressively assimilated within private minds, in an unending process of 'education'.

For Vygotsky, learning occurs in a contingent 'zone' where 'bridging' takes place between differing perspectives[212], 'scaffolded' by teachers, as guides interacting with learners[213]. Furthermore, individual development allows this zone to expand and be led outwards in particular ways[214]. The phenomenon grounds motivation for learners and teachers to participate in agentively-sourced, yet collaboratively-achieved intercommunication. Duties are assigned to the latter to identify modes of private thought, together with their present bounds. Planned, feasible, future extensions, successfully assimilable by the former, are based on these.

The relationship between inner and outer worlds of Sartrean and Piagetian formulations creates space for interactivity and reciprocity in authentic social relations. The one may contribute to shaping the other; stasis is absent. Continual reformulation and cultural reshaping take place through joint activity and especially, communication. Balances however, may be unequal. 'Political' power may come into play in determining the form and flow of particular communications.

Similar Vygotskian understandings are propounded in the libertarian work of Paulo Freire (1969)[215], and the language-based analysis of socio-cultural and political power by Norman Fairclough (1989)[216]. Writers such as these raise ethical concerns for axiological influences, attributable to constraints delineating any cultural environment, including those set by any chosen regime of assessment, moderation and evaluation.

However in criticising such approaches, Glaserfeld (1989) amongst others, has stressed that through mentally organising experience, subjects necessarily create personalised meaning. They attempt to avoid conflict through accommodating constraints on interaction, sourced in worlds external to themselves[217]. Individual shapings of language use are ever-present, since no two personal situations, temporal, cultural or linguistic, can be identical or replicated.

From ontological and epistemological perspectives on learning implicit in the foregoing, language is characterised as the key component of active, communicative interrelationships between 'self' and 'other'. It alters the evolution of otherwise unique relationships[218]. Within such constructivist views however, an established, cultural dimension, determining and defining norms, is highlighted as the referent for all 'acceptable', or 'standard' language use. Vygotsky represents coherent communicability as an attainment of individual consciousness, achieved through successful internalisation of shared social behaviour, and related to particular linguistic environments. Internalisation allows mental 'self-regulation' and the refinement of individual behaviour as a socially-communicable phenomenon.

In the genesis of consciousness, such thinking does not refute any claim to primacy for individual subjectivism. 'Meaning' in language is defined as privately-constructed, though adjusted in interactions with the social and linguistic world of 'other'. For any given individual, it is constrained within the bounds of the personally comprehensible. These bounds define 'zones of proximal development'. Adjustment to learning within such psycholinguistic zones is promoted by addressing problems of competition for meaning. Cognitive conflict is dissolved in a struggle between 'self'-constructed perceptions and the contrastingly dissonant propositions of 'other', resolved through harmonisation as chosen by the individual concerned.

The perspective has been developed with greater precision. Educational psychologists such as Bruner (1986, 1996, 1999), posit abilities in subjects to understand and respect the status of 'others' as alternate subjects, through active communication within social contexts of whatever configuration. For Bruner, learning is fundamentally linguistic and contextualised by cultures that establish viable language through social intercourse. These cultures publically legitimise themselves through the creation and operation of jointly-accepted institutions, within whose structures communication takes place. This allows individual intention to be expressed with meaningful, cultural "congruence" in the negotiated agreements that permit 'societies' to construct their characteristic meanings[219].

Hence, the Sartrean and Piagetian, inner 'self' is interconnected with a world outside, through meaning-making that links individuals to historical, socio-cultural and organisational, or institutional *milieux* in which they exist, or rather 'choose' to exist and express their individually-determined identities. (Self-chosen isolation from such *milieux* is always possible). Bruner's model of mind shapes a distinctive pedagogy and means for assessing learning of central relevance[220]. Yet attributing value to meaningful language, taken by Bruner as the fundamental operation of assessment, is also embedded within meaning-creating culture. The processes in play, albeit intersubjective and interactive, may be described and analysed either at a macrolevel of social and linguistic contextualisation, or at a microlevel, as in the **IBO**'s sociolinguistic assessment practice[221].

The position does not involve adopting and reiterating views such as those of Adorno. For Bruner, a postulated perspectival 'tenet' of ontology and epistemology permits individual and subjectively-sourced idiosyncracy in the creation and interpretation of meaning, within the situated context of "a culture's canonical ways of constructing reality"[222]. 'Culture' in this sense, is taken as in the sociology of Bourdieu (1991), to represent the operations of 'exchange systems', focalised and legitimised by their own institutional forms and by the 'symbolic apparatus' they employ as expression[223].

A more fully-developed position is succinctly reviewed in Bruner's recent work[224]. Learning is defined as contingent upon four key domains in the ontology of mind. For Bruner (1999), the acquisition of skill and knowledge allowing individuals to communicate with ‘other’ minds, is dependent upon recognition that learners are:

- *agentive:* that is, active in seeking out problems requiring resolution for mental equilibrium; interactive in engaging with environments surrounding them; selective in focussing the attention of consciousness upon elements deemed interesting and relevant in these environments; constructive in appropriating knowledge of ‘self’ in interaction with environments; and purposive in orienting all activity towards some goal, be it personally or socially chosen[225].
- *reflective:* that is, sense-making of acquired knowledge through matching (or mismatching) with privately-formulated and held hypotheses; capable thereby of interiorising the products of knowledge-acquisition and of reformulating their mental representation without further reference to environments from which they were acquired; and reflexive: that is, capable of turning the processes of internalised thought into the object of further thought.
- *collaborative:* that is, seeking out other minds for interaction through discourse; engaged with others in the solution of problems, the selection of focal points for attention, the construction of shared knowledge within a social arena, especially through talk; the achievement of common purposes, or goals both shaping and emerging from the construction of all social knowledge.
- *contextualised by culture:* that is, situated in cognitive ability within environments shaped by events of the past, present and imagination of the possible, to which they contribute

> collaboratively with others in a permanent, on-going evolution[226].

The development of constructivist and situationist views on mentation and inter-subjective interchange, has led to the devising by researchers such as Lave and Wenger (1991), of theories of cognition expressed in "apprenticeship models of learning". In these, selfhood is seen as expanding through interactions with experienced "masters" whose guidance in facilitating the acquisition of "mastery" by the "apprentice" learner serves as a socio-culturally, and institutionally-embedded framework for assessment and evaluation.

Moreover, in considering dynamic modifications to human mentation attributable to relational, and subsequently transformational dimensions in all language use, Bredo (1999) for example, emphasises both symbol-processing and situated cognition as significant, complementary models, integrating such theories of learning. For Bredo, individual minds are interactive, with knowledge and the representation of knowledge constructed through purposeful, situated, problem-solving activity based in symbol-processing and language[227]. Such activity is seen in a Deweyan sense as a 'transaction' between the individual and environments, whether socio-cultural or physical, changing both in the process[228]. Individuals and the environments in which they are situated are separate, but intimately and necessarily related through mental processing and acceptance that the 'reality' on which all symbols are based, is personally and internally representable. Language, as a socio-cultural tool permitting symbol-processing activity, is of central importance, even though any ultimate representation remains subject either to annihilation, or to transformation through individual, internal choice, and is always 'in the mind of the beholder'[229]. Whilst rooted in acknowledgement of the importance of Vygotskian ‘zones of proximal development’ for extending repertoires of personal mental representations as knowledge, Bredo views cognition as both ‘situated’ within definable contexts, and ‘constructed’. Such cognition is interactively modifiable and forever coming into being within ever-varying ‘cultures’, both shaping its forms of expression and themselves shaped by its further evolution.

For ‘success’ in assessment terms, Bredo requires an assumption of ‘task stability’, where problems to be solved are understood as simultaneously situated. First, such situations are within individual minds seeking to apply themselves to resolving given problems through volition, the adequate focussing of attention, and the selection of appropriate skills and knowledge for their resolution, whether successful or not. It is also within determinate, socio-cultural and temporal contexts in which task-designers seek in as unproblematic a fashion as possible, sensitively to design tasks that are recognisable and relevant. Through reference to shared, socio-culturally constructed knowledge, they should stimulate motivation to respond by being interesting and worthy of solution. Requirements for ‘recognisability’ thus become central to establishing authenticity in language use for assessment. They allow successful initiation of communicative, problem-solving activity, through appealing to ‘realism’ in assessment-task design and referring to materials sourced in ‘realia’. The latter are appropriated from relevant contexts situated in the socio-cultural, physical, temporal and linguistic environment forming both the context and goal of study, but not necessarily relating to the immediate assessment exercises in themselves. In this perspective, ‘failure’, as an assessment verdict may be defined as:

> "performance not understood in [the] light of potentially different, socially-organised interpretations of a situation."[230]

Within a perspective of situated learning, Lave and Wenger (1991) have theorised relationships between learner and assessed on the one hand, and tasks designers, standardisers, administrators, assessors, moderators and evaluators on the other, as *Legitimate Peripheral Participation*. Learning is viewed as gradual acquisitions of culturally-valued skills and knowledge, within socio-culturally contextualised relationships between novices, apprentices and expert 'masters' through whom it is continually referred to traditions transmitted to apprentices for the promotion and consolidation of new mastery, and subsequently assessed[231].

These perspectives are comprehensively summarised, updated and developed in the work of researchers such as Rogoff (1990, 1999). This extends the approaches of Bruner, Lave and Wenger, for whom pedagogy and learning are defined as a social, yet 'intersubjective' interrelationship of subjective minds and their partners, initially perceived through individually-selected processes of focussing attention, and then made social through a "joint establishment of focussing" of such attention between learners and 'teachers'[232]. Education is defined as the construction of communicative and dialectical relations, termed 'bridging' from one world-view, to another, and *vice versa*, through the Vygotskian and Brunerian activities of instruction in 'zones of proximal development' and 'scaffolding' through "guided participation". Such education reduces ambiguity, allows for effective, communicatively-based interaction, and the attainment of 'mastery' in any given knowledge area or skill of interest to the subjects concerned. Participation in education requires sharing and the joint structuring of efforts to solve mutually-accepted problems through "innate intersubjectivity" affording the taking of turns between subjects, and the focussing of attention by 'self' upon the intentions of 'others'[233]. This results in progressive transfers of responsibility in learning from teacher to learner. Ultimately for all, the appropriation and realisation of learning as a 'transformation of participation', is facilitated within a definable and purposive social tradition, as well as within the framework of unambiguously understood 'apprenticeship'. Significantly, these perspectives delineate and stress the joint importance of exotelic and endotelic purpose in all authentic learning activity[234].

To summarise, subjects both as 'self', and as 'object', or socially-perceived 'other', continually construct new relations of participation in common, shared, purposeful and goal-oriented activity. These relations allow for transformations of perceptions and understandings, and go beyond the mere reproduction of existing social mechanisms, which in Marxian terms, are achievable in material, economistic and historically-deterministic ways. For researchers such as Rogoff, and Lave and Wenger, this "apprenticeship learning", constructed through the process of "legitimate peripheral participation", allows recognition of necessary linkages between individual learners in their situation within defined socio-cultural, intersubjective and linguistic relationships. The latter are constructed through practice, requiring intentional engagement in the solving of dilemmas, individually identified as relevant and requiring resolution. Furthermore, it demands appropriate and voluntary focussings of attention. Thereby, it necessitates negotiations of joint meanings for the resolution of problems and the resultant satisfaction of previously-recognised 'needs'[235].

Such theoretical perspectives, linking individual intentionality in apprenticeship-learning as participation in social and linguistic relations with socio-cultural negotiation in assessing and evaluating the 'success' of this learning, render conceptualisations of authenticity for measuring and evaluating the qualitative coherence of interactive communication, both relevant and interesting to research.

Communicative Language Acquisition and Use

Developed conceptualisations of second language acquisition and use should be considered together with ontological, epistemological, axiological and cultural concerns such as these. All discrete languages are categorisable, describable and analysable, not merely as structured systems of sound and word formation, comprehension and production within a given framework, treated as independent of users, recorded, standardised and little-changing, but also as communication systems permitting the situated interlinking of 'self' with 'other' in a fundamental dialectic, forever creating, maintaining and modifying usage and meaning[236]. This linguistically-based, particularly-contextualised, cultural dialectic is reproducible in performance, and may be analysed for assessed and evaluated validation[237].

Sociolinguistic theories of culturally-contextualised, communicative language for teaching and learning have been developed contemporaneously with work in existential psychology, advancing learning theory. These are significant for Hymes (1971, 1974, 1977) who has theorised 'communicative competence' as acquiring Chomskian concepts of structural 'competence'[238], demonstrable in constructing communicatively-appropriate interlocutions between individuals and others[239]. Applied theory has been further developed by

Widdowson (1978) and Krashen (1981), amongst many others for practical evaluation purposes, but particularly so by Canale and Swain (1980), who use performance observations to infer assessed, theoretical levels of individual, linguistic competence[240].

Pedagogically-speaking, communicative theory has emerged from earlier approaches such as the structural, *Direct* and *Audio-Lingual* methods[241], popular in the 1950s and 1960s, which in turn had found favour as replacements for traditional, grammar/translation based learning that was orally non-communicative, relying almost exclusively on reading and translation for acquiring lexis, mastering grammatical rules and modelling in accurately reproduced writing, and being typified by dictation amongst other exercises. *Direct* and *Audio-Lingual* methods emphasised second language acquisition in behaviouristic ways, stressing as foundational, model-listening and speaking in drilled dialogues, with reading and writing developed subsequently. Communicative methods, evolving from the 1960s onwards, have retained this prioritisation of skill acquisition, but emphasised similarities with the acquisition of first languages, through lengthy, repeated exposure to target language, capturing listener attention with functional purpose, or otherwise and especially in 'authentic', real-time listening. Speaking is developed in situated, motivated negotiations of meaning with peers, or teachers as interlocutors. Reading and writing follow as before, as subsidiary skills aimed partly at recording and consolidating learning[242]. Teacher-interlocutors are 'facilitators' who elicit assessable performance. Hence as with the innovations of *Direct* and *Audio-Lingual* methods, communication takes place monolingually, wholly within the language to be learned.

For Krashen (1981) however, learning consolidates acquisition non-behaviouristically, through collaborative intercommunication, requiring peer and 'facilitator' interaction, with students and teachers selecting and organising 'curricula', often in cognitively-undemanding ways that refer to known ideas, experience, vocabulary and grammatical structure[243].

Combined with relatively simple use of cognitive capacity, skilled, functional knowledge permits sociolinguistically straightforward communication. Functional-notional, communicative curricula typically specify learning objectives and activities as naming objects and actions from a pre-defined nominal and verbal lexis, with their negative corollaries and set in elementary structures. These are applied to performance, simulating situationally-realistic and appropriately-contextualised projects for dialogue. The purpose of communicating is either intrinsic, or held as practical preparation for future application in similar encounters, assumed as likely. For Krashen, learners demonstrating soundly-acquired, communicatively-skilled knowledge may recognise new combinations of known lexis, phraseology or simple utterances and produce single words, short phrases or sentences with little subordination and without sophistication. Typically, these learners are 'ready' for 'meaningfully' communicative assessment within approximately four years of exposure to instruction in monolingual, classroom environments.

The **IBO**'s programme philosophy and design for *Group 2 Languages*[244] illustrates many of the features of Hymes' notion of 'communicative competence', set within general, idealised aims and practically-measurable objectives. They match Krashen's requirements for application, but have been further developed for authentic, self-defined, but situated usage with clearly-communicated purpose[245]. Communicative approaches are favoured, not so much as highly-structured, 'imposed' pedagogy, but rather as judicious eclecticism, encouraged as 'good', classroom-teaching practice. The development of **IBO** programmes, in particular from the 1960s onwards, has required changes to traditional pedagogies and curricula to suit new, more inclusive, heterogeneous, and in particular, international groups of learners. Hitherto, 'grammar/translation' or purely 'structural' methods, with their bilingually-based or behavioural emphasis on context-free pattern drilling and scant regard for communicative value in unpredictable, interactive, teacher and learner production, were seen as inappropriate, and all too often ineffective. Earlier teaching assumed that learners' motives unquestioningly included mastering target language in any and all aspects, final objectives often being notably literary in 'standard'. Fully-communicative, authentic approaches sought better to involve those with alternative linguistic experience, motives and interests in learning: 'usefulness' for practical purposes could still benchmark programme design and pedagogical method, even though learner input in determining purposes and assessing 'usefulness' had rarely been either investigated, or influential in securing satisfying curricula and pedagogy[246].

To summarise, contrast and develop further, communicative method in 'second' language learning stresses language as a complementary medium of communication, added to learners' existing, linguistic repertoire for social purposes in which learners have something to say or discover. At elementary levels, this embraces varying functions, ranging for example from presenting oneself in introductions, seeking information, expressing likes and dislikes, negotiating 'recognised' problems, and so forth, integrating these with socio-culturally embedded notions such as apologizing, asking for information or explanation, expressing opinion, *et cetera*[247]. Within this framework, classroom activities emphasise opportunities to use target language in communicative ways, personalising activity[248]. Concentration on meaning through message-creation or set-task completion, rather than on correctness of language-structure, pronunciation and vocabulary-choice, replicates modes of first language acquisition and allows 'second' language development as an alternative mode of communication. With target languages employed for classroom management and instruction, communicative approaches highlight 'naturalism' in acquisition, focus on perceived, individual 'needs', with ideal programming being Piagetian in structure, and claiming egocentricity as predominant in learner perspectives and interests. Student 'needs' include learned abilities to survive and cope in a variety of everyday situations, emphasising the acquisition of usable language, and exhorting learners to believe that they will indeed have contact with other peoples, prepared for the sociolinguistic and cultural realities to be discovered. In short, communicative approaches individualise and localise language, adapting it to supposed student motives for learning (as constrained by programme-designers), holding meaningful language to be thus more easily and securely retained[249].

The method requires primacy for listening and oral work. Contact time with target language is all-important, giving rise to more accurate command of structure and lexis, greater interpretative facility and more fluent expression. Making mistakes is 'natural' in this view of learning. Students using language purposefully, creatively and spontaneously are bound to make errors. Constant correction is unnecessary and often undesirable, being contained within a framework of communicative negotiations of meaning. (Here for example, error-correction may be required for clarification of obscure meaning).

Extensions of grammatical and lexical knowledge are subsumed within the method.

However, communicative pedagogy is not just limited to promoting aural and oral skill. Reading and writing are developed for increasing confidence in all domains of language use, though often, they are not proposed for their own intrinsic interest and value as determined by learners themselves, but rather as teacher-determined and directed support in consolidating mastery of spoken language for stably-recorded, 'objective' assessment. In using elements encountered in varied forms such as reading, summarising, debating and so forth, the manipulation of language becomes more competent.

Communicative approaches also notably emphasise recourse to 'authentic' resources, as those produced by target-language speakers, for their own, not-necessarily pedagogical purposes. They serve culturally to contextualise and support language-learning. In the classroom, 'authentic' texts partially substitute for direct communication with native-speakers. Typically therefore, reading material is extracted unmodified from newspaper and magazine articles, poems, manuals, guides, recipes, telephone directories, radio, television and cinematic material and very often news bulletins, discussion programmes and so forth, for various exploitation in building and consolidating all language skills[250]. Krashen summarises communicative theory thus:

> "What theory implies, quite simply, is that language acquisition, first or second, occurs when comprehension of real messages occurs, and when the acquirer is not 'on the defensive'. [.....] Language acquisition does not require extensive use of conscious grammatical rules, and does not require tedious drill. It does not occur overnight, however.

> Real language acquisition develops slowly, and speaking skills emerge significantly later than listening skills, even when conditions are perfect. The best methods are therefore those that supply 'comprehensible input' in low-anxiety situations, containing messages that students really want to hear. These methods do not force early production in the second language, but allow students to produce when they are 'ready', recognizing that improvement comes from supplying communicative and comprehensible input, and not from forcing and correcting production."[251]

The Identification of Components of Authentic Language Use

To return briefly to Sartre, authentic, communicative interaction creates coherence and validity in all social relations. The processes are dynamic, motivating further intention to interact and construct identity. A continual project for becoming emerges from needs to resolve intrinsic tensions posed in conflictual, non-negotiated choice proposed by others. All learning activity is necessarily socially-embedded, problem-solving and constructive. In terms summarised by Edwards and Mercer (1987)[252], learners are inducted into an established, ready-made culture, whilst interacting as aware and autonomous participants within that selfsame culture, as it is continually in the making. The interactive medium of language use expresses, transmits and transforms such culture as one of its fundaments. For language learning, and in particular, for 'second' language learning, typical interactions may be defined (albeit in simplified summary), by specific qualities, identified for example by Van Lier (1996)[253].

In conceptualising authentic language use, Van Lier views authenticity as the third of a 'triad'[254] of interlinking concepts defining communicative interaction in any given code (and more particularly in any given language pedagogy and curriculum[255]). The set includes "awareness" and "autonomy" as preconditions for the construction, or attainment of authenticity, on which this third state constantly depends. Nevertheless, interactions modifying the form and status of prior concepts are ever-present possibilities. For individual minds, according to Van Lier:

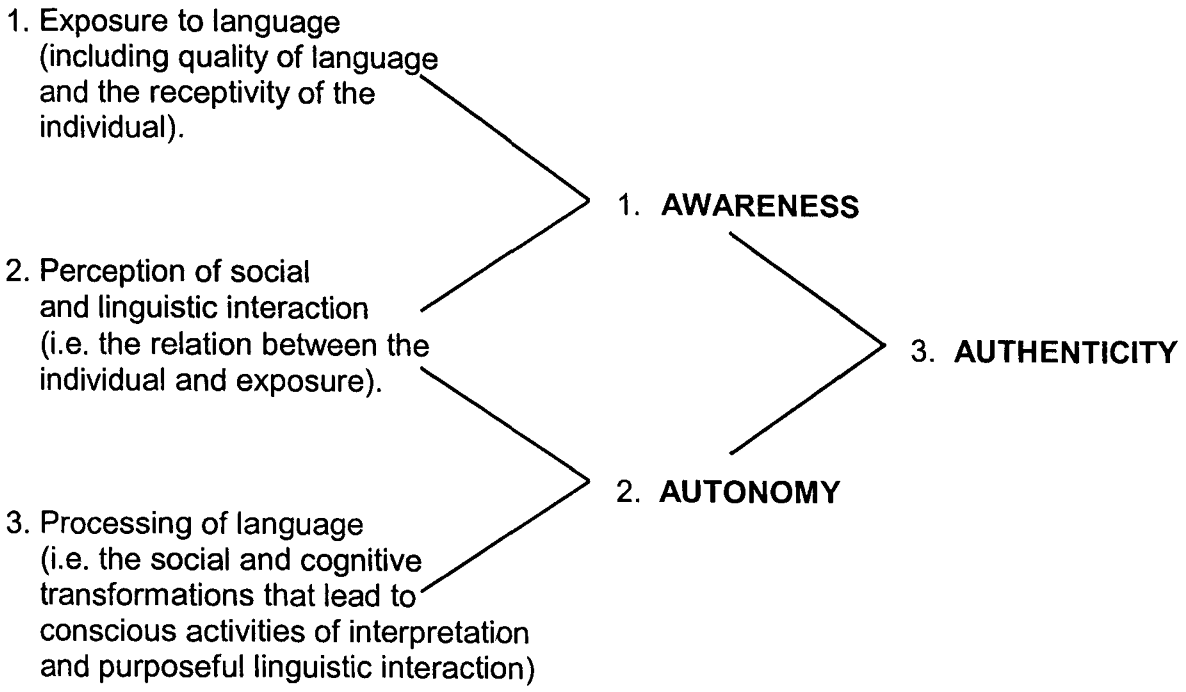
> "It might be argued that authenticity is the natural result of awareness and autonomy, and at the same time, that authenticity leads to increased awareness and autonomy. In other words, if you 'know what you are doing', and if you are 'responsible for your own actions', then you are 'being authentic'. [.....] If awareness is firstness, autonomy (one's stance *vis-à-vis* others and the object world) is secondness, and authenticity is the interpretation of that which unites awareness and autonomy."[256]

From complex argument, findings are summarised by a typology founding an assessment-instrument design for identifying and measuring features of authentic language-production[257].

In this perspective, communicative language pedagogy and learning, defining authentic relations between teachers and learners, may form a further triad, interlinking "curricular", "pragmatic" and "personal authenticity". This is intimately related to a triadic model, illustrating all authentic expression, proposed and diagrammatically represented by Van Lier, as follows, (with all lines representing possibilities of interaction):

Figure 3.1

A triadic representation of interrelations between awareness, autonomy and authenticity



1. Exposure to language (including quality of language and the receptivity of the individual).

2. Perception of social and linguistic interaction (i.e. the relation between the individual and exposure).

3. Processing of language (i.e. the social and cognitive transformations that lead to conscious activities of interpretation and purposeful linguistic interaction)

(Adapted from Van Lier[258])

In this diagram, features identified as "curricular", "pragmatic" and "personal" authenticity interrelate with 'exposure' to language, 'perception' of social and linguistic interaction and the 'processing' of language units. Understanding this interlinkage is easier when the three former concepts are broken down into a supplementary triad, serving to isolate discrete components of authentic expression[259]. That is, "curricular authenticity" resides in an individual's possibilities for *using* and *creating* language, after exposure to models *found*, or *received* by the individual from the linguistic environment. "Pragmatic authenticity" relates to individual *purpose* in public language-production, and hence to physical, temporal and socio-cultural *contexts* within which linguistic *interactions* take place. "Personal authenticity" subsequently emerges from resultant, linguistic processing, establishing ontological, or *existential* status for individuals committing themselves to the interchanges that take place through *intrinsically-motivated*, endotelic choice. Integrating committed participation in such interchange with inner-sourced, purposeful, or goal-oriented motivation results in what the educational psychologist Csikszentmihalyi (1990) has termed an *autotelic* personality[260].

As Van Lier claims, these categories may be better understood as criteria supplying evidence for *pragmatic authentication*[261], a concept that is further defined, discretely categorised and discussed in reporting on research methods[262].

Authenticity and the Measurement of Linguistic Attainment

In further review of relevant concepts, the following chapter focuses on a third dimension of authenticity. Assessment theory and practice for evaluating communicative pedagogy and its associated learning, are related to contextualised task-setting for stimulating authentic language use through performance[263]. In this respect, general understandings, developments and use of what is often, traditionally labelled as 'authentic assessment' *per se*, are reviewed.

Following extensive discussion amongst specialist commentators such as Bachman, (1990), Bachman and Palmer (1996), Gipps (1994), McNamara (1996) and others, authenticity in assessment is understood as the design and administration of testing situations that minimise to the greatest degree feasible, the general constraints on 'self'-expression in socio-culturally interactive and meaningful ways, within cognitively-situated contexts.

Hence, reviewing the literature of authentic assessment should illuminate rationales for considering validity and reliability as key issues in problematising assessment and evaluation. With 'high stakes' testing[264], such investigation should reveal relationships amongst alternative measurements and quantifications, attributing value by numerical score. Considerations of equity evidently, also come to the fore.

CHAPTER FOUR

THE LITERATURE OF ASSESSMENT

The Design and Standardisation of Communicative and Assessable Tasks

With sociocultural and communicative, rather than psychological, or psycholinguistic definition of language use, test constructs are of primary theoretical concern for the **IBO**[265]. McNamara (2000) cautions, however:

> "The term test construct refers to those aspects of knowledge or skill possessed by the candidate which are being measured. [.....] Defining the test construct involves being clear about what knowledge of language consists of, and how that knowledge is deployed in actual performance (language use). Understanding what view the test takes of language use in the criterion is necessary for determining the link between test and criterion in performance testing"[266]

Through emphasising communicative and authentic expression in designing, standardising, assessing, moderating and evaluating parameters for formal assessment, **IBO** definitions and practice show lesser concern for measuring control and mastery of pre-defined features for given language-systems than has been traditional[267]. Conventional approaches refer to the constructs of structural linguistic theory, favouring discrete-point measurement of isolatable elements of discourse, traditionally categorised by grammatical classifications, or through identifications of uncontested, stable meanings in lexis, often assessed in de-contextualised fashion in separation from syntactic structure, and occasionally with recourse to translation into another language[268]. For testing within psychometric paradigms, such constructs are most usually designed to enhance reliability in evaluation, illustrating the belief that language use and performance are stable and available for relatively unproblematic, measurable representation through objective, non-dialogic, non-interpersonal and non-pragmatic methods. Assessment exercises consequently lead to the evaluation of linguistic production, predominantly measured in fixed, time-independent and at least partially non-interactive forms, stabilised and constrained within pre-defined corpora of possibilities. These are provided more readily by written, rather than oral use of any given language. Hence, psychometric measurement is strongly biased towards the study of language through comprehension, rather than production. Psycholinguists such as Garnham (1985) have usefully summarised the reasons[269].

Psychometrics and the Approach of Psycholinguistics

Psycholinguistic theory places little emphasis on language either as integrated knowledge and skill, displayed in interactive, communicative performance, or as the heuristic linkage of language-producers with audience and readership, assessors and evaluators, whether such performance be contextually authentic or not. It highlights the practicality, effectiveness and credibility of psychometrics for 'objective', positivistic assessment and evaluation, considered feasible as independent, unreactive activities in their own right, and achievable through the greatest possible reduction of non-predictable variables in interactions, since these create 'noise', or distortion in taking measurements. Psychometric systems are considered capable of producing valid results with high levels of statistical reliability[270]. Psycholinguists designing them propose the elimination of assessor and evaluator subjectivities as invalid and unreliable influences in categorising and measuring language use. Instead, recourse to test-item scorers (or machines), mechanistically assessing language-production through applying pre-determined, discrete, norm-referenced constructs, and aggregating responses scored as 'correct', or 'incorrect' within a numerical grading-system, is considered sound method. In particular, the psycholinguistic method of psychometrics is characterised by functionally-determinant categorisations of discrete language skills, demonstrating knowledge through listening, speaking, reading and writing, with weighting for higher levels of achievement placed on proficiency in producing richly-varied, clear, elegant, 'standardised' language in written forms[271].

Furthermore, in contradistinction to communicative approaches in measuring language comprehension and production, (with their concerns for authenticity), psychometric method assumes identity as 'self' to be irrelevant in assessing production. The passive, mentally-internal skills of comprehension are measured in alleged isolation from the active skills of speaking and writing, or indeed of focussed, purposeful listening and reading[272].

The approach implies a view of mind and learning that considers individual test subjects as asocial and decontextualised from particular environments within which test performances are completed. Comprehension may be signalled by a variety of means, often either non-linguistic, or dependent upon some form of reference to a second, extrinsic, securely-mastered and untested, 'common' or 'standard' language, assumedly shared by candidate and assessor, and accessed through paraphrase, interpreting and translation. Psychometric assessments and evaluations separate language from interactive projects for communication. Competence is measured as quantity in mastering a stable corpus of given linguistic knowledge, independent of individual producers, and benchmarking determinations of quality in candidate productions for assessment. Psychometric theory confidently advocates testing regimes that place high value on establishing a minimally-contestable rating reliability.

Communicative Language Use in Assessment and Evaluation

In attempts to overcome problems for assessing and evaluating authentic, performance-based and situated uses of language, Oller from the 1970s onwards, theorised a more integrative approach to test-design, standardisation, assessment and evaluation[273]. On the one hand, test constructs should concern linguistic processing in temporal, often 'real-time' contexts, as in aural reception for comprehension and oral production. Hence, oral and listening assessments, discrete or integrated, gained value as significant components of valid, testing programmes. On the other, such constructs should focus on sampled simulations of practical usage, through facilitating the evaluation of aspects defined as:

> "the ability to integrate grammatical, lexical, contextual and pragmatic knowledge in test performance"[274]

In measuring knowledge and mastery of written language, including written indications of listening and reading comprehension, reliability for

Oller (1979), is achievable through careful test-design, including candidate-productions in assumedly controllable and contextualised situations, typified for example by cloze-testing. This method appears particularly reliable when conjoined with multiple-choice, target-language options for possible answers: a technique and procedure claimed appropriate for measuring competence in comprehension. It requires the successful recognition of established grammatical patternings and stable lexical usage, by candidates whose roles, purposes and abilities are defined through selecting 'correct', substitutional 'answers' for unambiguous scoring. The validity and reliability of such assessment-task creation, Oller proposes, may be further enhanced through comprehensive application of statistically-derived controls at the design and standardisation stages[275].

However, such approaches are also evidently appropriate for understanding linguistically-acquired or constructed cognition as time-independent processings of symbols, fixed unchangingly in written formats. They disregard the temporally-changing, socially-contextualised nature of necessarily interactive, communicative language use, typified by the close integration of aural comprehension with oral production, in the dialogue of authentic conversations and their like. In this model, listening comprehension is measured through modes of reading (the tasks set), and writing (responses to the tasks set). It too is critically dependent upon basic competence in these distinct skills.

The conceptualisation of 'communicative competence' by Hymes[276] as a sociolinguistically-measurable construct for language-test design and standardisation, duly integrates the temporal and socio-cultural situation of all linguistic intercourse, including influences that each situation brings to bear upon receiving and producing language. Thus, Hymes' sociolinguistic research pleads for language-testing to move away from structuralist psychometrics, since this discipline is anchored in given linguistic certainties of contestable validity as universals, and privileges the skills of reading and writing. Instead, attention should be paid to designing scenarios and role-play simulations for examination candidates who are recorded performing extended acts of communication within supposedly 'real', or in this way, socio-culturally (if not individually) 'authentic' contexts and identities for each 'performer' under test. Representational realism, simulating specified socio-cultural and functional roles that are imposed upon test candidates through test rubrics, should 'motivate', 'appropriately' initiate, and thus facilitate assessable communication. The gains in validity are claimed reliable on the basis of commonality of task for all candidates in any given testing programme. The principal theoretical concern for test designers and standardisers thus becomes the development of stable categories for measuring performance within known role-identities, contextually-predefined in requirements for displaying linguistic and communicative competence.

From Hymes' work, applied theorists such as Canale and Swain (1980)[277] have categorised language for discrete assessment and evaluation as evidence for:

- *grammatical competence*, demonstrating knowledge of formal features of language structure and lexis;
- *sociolinguistic competence*, demonstrating knowledge of 'appropriateness' in discourse, in recognisably patterned, targetted social situations shared between users of the given language;

- *strategic competence*, demonstrating knowledge and ability to adapt linguistic performance to unpredictable or imperfect occurrences in linguistic interchange;
- and *discourse competence*, demonstrating knowledge and ability to maintain linguistic reception and production coherently over extended periods of time[278].

Similar, though more detailed itemisations of competence have been reformulated by others such as Bachman (1990)[279]. However, besides specifying categories in increasingly more finely-grained, discrete identifications of components of linguistic knowledge and skill, these researchers have highlighted increasing awareness of the complex, problematic nature of conceptualising linguistic phenomena as strategic and discourse competencies. That is, interactive, and thereby authentic language-production partly rests upon fluid, temporally-based and ever-changing, dialectically-communicative relationships between two or more speaker-listeners or reader-writers, sharing a linguistic code. This is most evident in interlocution, though may also be true for the demonstration of comprehension, and in the production of writing for any given readership, in response to any given stimulus. The design and standardisation of relevant tasks and rubrics favouring authentic expression in forms permitting valid, reliable and equitable assessment, leading in turn to quantifiable and explicitly justifiable, consistent evaluation, are therefore far from simple.

McNamara summarises the dilemma in developing testing theory:

> "[.....] the approach to thinking about communicative language ability in terms of discrete components leaves us with aspects of language analysed out as distinct and

> unrelated. There is still therefore the problem, which models of communicative competence were designed to resolve, of how to account for the way the different aspects act upon each other in actual communication. Paradoxically, as models of communicative competence become more analytic, so they take us back to the problems of discrete-point testing usually associated with testing of form alone.
>
> [.....] A further issue involves the implications for test validity of interpreting test performance, for example on a speaking test, in terms of only one of the participants, the candidate. Clearly, many others than the candidate affect the chances of the candidate achieving a successful score for the performance. These will include those who frame the opportunity for the performance at the test design stage; those with whom the candidate interacts; those who rate the performance; and those responsible for designing and managing the rating procedure. Instead of focusing on the candidate in isolation, the candidate's performance needs to be seen and evaluated as part of a joint construction by a number of participants, including interlocutors, test designers, and raters."[280]

The problems beg alternative approaches to their resolution, with the present investigation of authenticity as categorisable in graduated descriptions and applicable in measuring quality in authentic language use, proposed for exploration in this sense.

Criterion-Referencing for Measuring Language Use

Before considering specific literature on performance and the authentic assessment of language-production *per se*, review of criterion-referencing theory is pertinent. The **IBO** adopts explicit criteria for benchmarking authentic and communicative usage for *Diploma Programme* internal assessments and examinations[281].

Summarising understandings developed to date, McNamara defines criterion-referenced assessment as:

> "an approach to measurement in which performances are compared to one or more descriptions of minimally-adequate performance at a given level."[282]

'Criterion' is defined as:

> "1. The *domain* [author's emphasis] of behaviour relevant to test design.
> 2. An aspect of performance which is evaluated in test scoring, e.g. fluency, accuracy, etc."[283]

Gipps (1994)[284] on the other hand, contrasts "criterion-referenced assessment" with "performance assessment", and distinguishes from this larger, latter category, a subset of "authentic assessment". For Gipps, criterion-referencing after Glaser (1963), is defined in opposition to 'norm-referencing', as:

> "Measures which assess student achievement in terms of a criterion standard [and] thus provide information as to the

> degree of competence attained by a particular student which is independent of reference to the performance of others."[285]

In contrast, performance assessment is understood as:

> "[an] aim to model the *real* [author's italics] learning activities that we wish students to engage with, oral and written communication skills, problem-solving activities, etc., rather than to fragment them, as do multiple-choice tests; the aim is that the assessments do not distort teaching."[286]

However, for Gipps, this purposive definition is further contrasted with a commonly-termed, 'authentic assessment'. Under the latter, clear intents are to minimise undesirable effects in 'washback' that threaten the validity of assessment constructs requiring authentic language use, and accompany any transparent, published evaluation system, as test-derived influences on choice of teaching and learning styles and content[287]. In her words:

> "authentic assessment is performance assessment carried out in an authentic context, i.e. it is produced in the classroom as part of normal work rather than as a specific task for assessment. While not all performance assessments are authentic, it is difficult to imagine an authentic assessment that would not also be a performance assessment."[288]

Gipps gives practical examples, citing amongst others, the production and collating of sampled assessment evidence in representative portfolios[289]. Referring to Meyer (1992), who in these contexts, requires assessors to specify all elements contributing to 'authenticity', Gipps lists:

> "the stimulus; task complexity; locus of control; motivation; spontaneity; resources; conditions; criteria; standards; and consequences."[290]

The key distinction between McNamara and Gipps *et al,* lies in contrasting unambiguous specifications, or the absence of such specifications, for interactively involving interlocutors, or 'co-performers', in assessable, linguistic performance. Given authentic assessment procedures, the resultant production is compared with descriptions representing norm-independent criteria. Explicit and detailed contextualisations for all assessed performance are also provided for due consideration, even though for test-validity, reliability and equity, the control and adjustment of these conditions may remain unclear.

The distinctions may prove significant[291], since assessment providers must first develop criterion-categories forming discrete elements of structure in any system of assessment and evaluation. With **IBO** *Internal Assessment*, aspects of authentic assessment as defined by Gipps, provide guidelines for year-long collections of evidence[292]. In all such formal assessment, whether independently moderated by the **IBO** or not, criteria are specified in the three domains of *Task and Message, Interaction* and *Language*[293].

As recounted, the **IBO** publishes no explicit rationale for such tripartite categorisation, despite its apparently Hallidayan origin[294]. Nonetheless, it can be understood from the organisation's documentation that categorisations of assessment criteria are eclectic, referring to various theories integrated into a whole for formal assessment purposes. Certain categories appear to correlate with theoretical constructions from structuralist linguistics, recognising an authoritative corpus of standardised language-knowledge. Under criteria described for *Language*, there is explicit reference to *Accuracy* in speaking and *Grammatical Accuracy, Vocabulary* and *Intelligibility* in writing[295]. Related, psycholinguistic concerns are reflected in identifications of psychological and personality-based 'skills', under criteria for *Fluency, Coherence, Interaction* and *Comprehension*, as defined for oral *Interaction* and written *Presentation*[296]. The concerns of sociolinguistic theory are evident in remaining categorisations. Indeed, these often overlap as requirements for assessing performance in sociolinguistic fashion, scattered through many of the specifications. Examples are constructs of "effective[ness]", "comprehensive[ness]", "relevance", "appropriacy" in register, style and content, whether lexical or ideational, "cohesiveness", "liveliness", "initiative" in language-production, the "fluency of pronunciation and intonation", the generation of "interest" for the interlocutor or reader, and so forth[297].

In identifying and categorising assessment and evaluation criteria, McNamara (2000) emphasises the necessity, first, to situate constraints under which test providers must operate as administrators[298]. These influence choice of test method, rubric and content, and are of particular relevance to international organisations providing 'high-stakes' assessments and evaluations, such as the **IBO**.

In considering test methods framing the production of oral or written language, and the formats required for candidate responses, McNamara further reminds us that authenticity is significant in either of its broad, twin conceptualisations: as constructs deriving from naïve understandings of representational realism, or as references to situated participation in interactive, communicative dialogue between 'self' and 'other'.

In the former instance, assessment and evaluation criteria will relate to domains representing sets of practical, 'real-world' tasks and requiring the identification and description of pre-defined, recognisably 'realistic', simulation-roles for candidates. They are representationally related to benchmarks derived from examples of 'real-life' performances, assumedly recorded in 'real' situations within 'real' time. The candidate is presumed to recognise and accept such 'realism', thereby 'suspending any disbelief', being motivated to participate either through a desire to prepare for future, expert and 'real' performance, or simply to perform well in a test situation.

In the latter instance, where authenticity promotes interactive, communicative, yet situated dialogue, the point of reference is "a theory of the components of knowledge and ability that underlie performance"[299]. This assumes candidate interest and intrinsic motivation to participate in such performance, not only for assessment purposes, but also for its own sake.

For any given test design, such concerns conflate problems of method with specifications of rubrics and content[300]. A designer and standardiser's dilemma can thus be summarised as a need for compromise. Tensions to be held in balance arise from two mutually-incompatible demands. Hence, for McNamara:

> "On the one hand it is desirable to replicate, as far as is possible in the test-setting, the conditions under which engagement with communicative content is done in the criterion-setting, so that inferences from the test performance to likely future performance in the criterion can be as direct as possible. On the other hand, it is necessary to have a procedure that is fair to all candidates, and elicits a scorable performance, even if this means involving the candidates in somewhat artificial behaviour."[301]

In such procedural matters, intervening demands for reliability from administrative orders emphasise deeper aspects of constraint. However, these are created for validly heuristic purpose and are independent of candidate choice of participation through motivated, authentic expression. They link test-designer, standardiser, test-administrator, and all individual candidates completing the relevant test, to assessment and evaluation procedures for the ensuing productions, according to institutionally-determined criterion-categorisations of discrete performance levels. The characteristics of any language willingly produced for assessment purposes by the actors concerned, and situated by phyṣical location at the points in time in which their relationships are framed, are thus intrinsically socio-cultural in their fundaments. For commentators such as Bourdieu (1991)[302] and Fairclough (1989)[303], they are thereby deeply 'political' as well[304].

With this type of assessment, **IBO** specification of criterion-descriptors for tasks of differing level, and the control of procedures by which these tasks are 'correctly' completed, imply issues in an institutional agenda requiring further investigation. Through subscribing as 'clients' and 'consumers' to **IBO** programmes, teachers and students alike adopt constraints that at once determine a categorical and hierarchical status for given, 'second' languages (as 'native', 'near-native', 'highly competent', 'foreign' and 'beginner')[305]. 'Clients' choose and distinctively 'encode' language for production at relevant levels, guided by the form and content of discrete programmes. They may recognise displays of semiotic 'power' for successful 'decoding' and respectful observance in situations of inauthenticity, if the award of an institutionally, often socio-culturally, and academically prestigious 'result' is desired. Such 'inauthenticity', deriving from specifically socio-cultural and temporal contexts in which task-setting, responses, assessment and the evaluation of such responses all take place, ensures that the 'dialogue' and 'dialectic' of language use in any given exercise cannot be fully, hence authentically, interactive and communicative.

This is particularly so in written production and its assessment, since under typical circumstances, engagements in chosen activities can neither contemporaneously, nor fully authentically link writer with reader (except perhaps in the modern-day use of electronic media formats). For the **IBO** indeed, typical tasks are discrete, and their composition will normally be separated from reading, assessment and evaluation by intervals of time of some weeks in duration. Meaning may thus not fully emerge in authentic forms as socio-cultural and linguistic construction, conjointly achieved through the exchange of producers and receivers, operating 'freely' within a 'linguistic market' – a market furthermore, whose 'rules of production' are determined at any given time, by an institution in its public role as arbiter and evaluator of the resultant products. Indeed, the negotiation of meaning and of boundaries for constraints within which meaning is produced can only be altered 'after the event' of assessment, on reception of candidate and 'client', or 'consumer' input within such a 'market'.

In any specification of language 'level' and of qualitative, criterion-referenced categorisations of language use for assessment and evaluation at given thresholds, the dilemma raised by commentators such as Bourdieu and Fairclough has been usefully summarised by Sanderson (1997) as a representation of the core, ontological and epistemological antagonism between 'culture' and 'subjectivity' in the discourse of assessment[306].

Sanderson defines the central problem of assessment as one of validity in "tension", inherent in its character as:

> "an individual act of judgement on the one hand, and on the other as a process which is profoundly cultural, a tension accentuated by the dichotomy between ideologies which hold knowledge to be objective and monolithic, and those which believe knowledge to be contingent, relative and plural."[307]

Bachman and Cohen (1998) amongst others[308], have referred to the emergence of such dichotomies in theoretical approaches typifying certain discrete (for Bachman, explicitly too discrete) concerns of researchers in various fields of linguistics. Such researchers have significantly influenced test conceptualisation and design, and have produced major tensions at the 'interface' (the term is Bachman's) of *Second Language Acquisition* and *Language Testing* theories.

For Bachman and Cohen[309], second language acquisition research has traditionally and predominantly emphasised issues in categorising, identifying, selecting, describing and analysing evidential data for learning, with subsequent induction of theory. Most often, this has been by qualitative methods imposing a paradigm of longitudinal, or diachronic views of what they term "interlanguage development". Its objectives are "to describe how second language acquisition takes place, and to explain why [such acquisition] takes place"[310].

In contrast, language-testing research has been typified by synchronic concerns for point-in-time sampling of language comprehension and production, referenced to given norms, or accepted 'standards' of language use. Its approach has often been influenced by structuralist concerns for language description within discrete categories of production. Classically, these have been conceptualised as language 'ability', psychometrically-measurable by reference to notions such as those for grammatical and lexical competence (an overarching concept covering the display of 'range', 'accuracy', 'variety', 'complexity', 'appropriateness to context', and so forth). The approach is often inevitably quantitative, attempting to:

> "develop and empirically validate a theory of language test performance that will describe and explain variations [.....], and [.....] demonstrate the ways in which [such] performance corresponds to non-test language use." [311]

Both strands of research interest illuminate the categorisation and designation of criteria in criterion-referencing schemes, as proposed by the **IBO**. They stress unresolved tensions distorting assessment and evaluation as pure measurements of authentic language use, though neither considers the significance of learner and candidate motivation for communicating through target language. Indeed, in concluding the chapter from which quotation has been made, Bachman and Cohen propose "directions for future research" requiring clearer discussion of issues for "characterising authenticity and the nature of language use tasks"[312]. Not enough seems to be known of the effects and implications of authenticity as volition to express and develop 'self' through the communications in question.

The Standardisation of Examination Tasks

Given emphatically socio-cultural, rather than structural or psycholinguistic definitions of language and its use in **IBO** documentation, theoretical discussion of standardisation highlights the importance of careful specification of test construct[313].

As defined and discussed by commentators such as McNamara, specifications of criteria serve as primary determinants, setting parameters in the design of assessment-tasks, and establishing reference points for standardisation across different administrations of common examinations at a single level. In this respect, theories of test performance based on structural linguistics and emphasising the discrete, predominantly psychometric measurement of mastery of pre-defined elements of given language systems, assume that sound, incontestable categorisations can indeed be formed[314]. These permit functional definitions for delineating clear boundaries in any given programme range and for the **IBO**, at *Higher* or *Standard Levels,* within *A1, A2, B* or *Ab Initio* categorisations of language knowledge and skill.

Despite evident problems of viability in categorising all-embracing, unitary conceptualisations such as those of authentic language use and 'familiarisation', the consequently enhanced possibilities for 'washback', implied by standardisation, influence choices of classroom curriculum, of teaching and learning styles. These effects arise when known, common criteria are regularly published for each level and grouping of languages, and for standardising differing assessment rubrics and tasks[315]. They also bear directly on authenticity as the promotion of time-dependent, individual, and hence irreproducible performances of listeners and speakers, writers and readers, culturally and socio-linguistically situated in communicative interaction and interchange.

As the term itself implies, the concept of standardisation is problematic. For the **IBO** programmes considered, it emphasises a necessarily discrete differentiation and categorisation of production and performance as 'standards', consistent for a single language and level, and graduated in a hierarchy of requirements. As such, standardisation necessitates the adoption, either explicit or implicit, of concepts of 'stability' at coherently-defined 'thresholds', as defined for example by Van Ek (1975, 1976), and Van Ek and Trim, (1991, 1996) for the **Council of Europe**[316]. Within each categorisation, authenticity becomes significant and potentially difficult to define, since any single set of concepts intentionally promoting authentic language use may be taken as broadly applicable to any level and grouping. In themselves, conceptualisations of authentic language use can hardly be determined by, or dependent upon defining discrete levels of language proficiency[317].

Hence standardisation as norm-referencing process, whether defined or not, presents conceptual difficulties for the design of any assessment and evaluation attempting to ensure full respect for authentic language use[318].

Categorisations of Language-Performance for Evaluation Purposes

In this instance, theoretical problems for review concern the specification and description of practical, valid and reliable criteria for evaluating evidence of assessable performance. McNamara (1996) warns against conflating these with criteria for task-design. By doing so, confusions in conceptualising authenticity either as representational realism, or as situated, communicative, interactive, linguistic interchange, are exacerbated. The following summarises the problem through posing a key question:

> "There is an ambiguity here: is the performance of [.....] tasks in the test situation 'valued in its own right', or are the tests in the real world valued in their own right?"[319]

For the **IBO**, overall performance descriptors, differentiated at graduated levels, are derived from key assessment criteria. These are specified, according to **IBO** 'espoused' theory, without reference to considerations of task-setting, or standardisation for evaluation. Evaluation descriptors should facilitate valid, reliable and equitable transformations of assessments into quantitative scorings, determined through matching samples of assessed language-production to appropriately-aggregated categories of qualitative description. Individual performances may thus be justifiably labelled.

The problems concern evaluation as further transformations of qualitative value into equivalent, numerical representations, expressed as grades. Transparency and consistency with programme philosophy, aims and objectives come into the reckoning, since these inevitably, are particular to any system-design. Linkages between seemingly distinct measurements for differing assessment and evaluation purposes, with varying instruments and in discrete contexts ranging from fully informal, 'authentic', or performance-based, continuous assessment, to fully formal examination, requiring individual responses to fixed tasks at a determined date and within a determined time-allocation, are all potentially difficult to establish in their own right.

As Bachman and Cohen (1998) show, the interface between second-language acquisition and language-testing research is revealed through statements of specific purpose for all assessment. The principal ethical rationale of evaluation is to indicate improvements in advancing learning and performance. These authors' research thereby stresses assessment value as qualitative description rather than numerical representation.

In accounting for 'washback', the emphasis points to possible distortion in the authenticity of evaluated language-production. Potentially significant 'washback' effects accompany all consultations of assessment criteria and systematic evaluation processes, published to promote institutional transparency and pedagogical familiarisation. For assessment, as Bachman and Cohen suggest[320], promoting authentic language use is assumed to stimulate the development of classroom curricula and strategies for encouraging and enhancing more successful, authentic communication. In a virtuous cycle, this leads in turn to improved performance.

Anachronistically, yet as if in reply to McNamara's previously quoted conundrum, Gipps amongst many others, stresses the importance of evaluation through 'authentic', criterion-referenced, performance assessment[321]. This mode, favouring authentic teaching and learning, offers viable alternatives to behavioural, structural language-drilling, often encountered and reported by researchers as preferred, classroom activity for preparing student success in 'high-stakes' examination, under the influence of familiarisation and 'washback'. Task-responses should be valued (and evaluated) in themselves, over and above any assessed matching that contrasts simulations with representations of a pre-determined and known external world to which behavioural understandings refer. In this, a key theoretical assumption is that all processes of assessment and evaluation should be explicit and transparent, thus aiding the non-behaviouristic, 'authentic' processes of communicative teaching and learning, with these appropriately socially-situated and involving meaningful, linguistic exchange that is integrally dialectical and interactive[322]. Indeed as Gipps hypothesises, the negative effects of 'washback' on authentic language use may be curtailed through continuous assessment, based on regular, structured sampling of portfolios of recorded, classroom interactions.

The literature suggests that 'high-stakes' programmes encourage behaviouristic approaches to pedagogy, if not indeed effective learning. Through publication for transparency, repetitions over time, recommended training for familiarisation and the harmonisation of teacher understandings with **IBO** philosophy and aims, formal assessment 'standards' and demands become increasingly stable and 'accepted' as objectives, if not aims, for teaching and learning. The consequences nevertheless, involve questions of curriculum and learning, rather than determine assessment and evaluation design and its applications. Thus in this research, such considerations have not been taken as central[323].

The Role of Examiner Training and Moderation

However, given **IBO** requirements and recommendations for training *Internal Assessment Moderators, Assistant Examiners,* and teacher-*Internal Assessors* through designated **IBO** workshop-training sessions, the conclusions of Gipps, summarised with references to Linn and Dunbar *et al.* (1991), Shavelson *et al.* (1992), and Linn (1993a), cannot be wholly ignored in any discussion of 'washback' in assessment and evaluation. Gipps claims that:

> "the weight of evidence reviewed by Linn and Dunbar [.....] indicates that score-reliability is generally low, lower than rater-reliability and more resistant to being raised than is [inter-rater reliability] through training, etc. The evidence is that performance on performance assessment-tasks is highly task-specific; that is, performance on different tasks from the same domain, or on tasks that appear to be similar, will only be moderately related. The actual task set leads to variability in performance; the method of assessment (observation, notebook, computer simulation) also affects measured performance, since each method provides different insights into what students know and can do [.....]. Increasing the number of tasks in an assessment tends to increase the score reliability more than does increasing the number of raters, and Linn [.....] advocates increasing the number of tasks to enhance generalisability."[324]

Given **IBO** designs for assessing and evaluating language-production, the requirement may appear desirable, though begging questions of status in candidate choice of task, of comparability in choice across alternative tasks, and of 'compensation' in determining equivalences between different responses, aggregated as totalised scores[325]. For enhancing reliability on which generalisations must be based, increasing assessment tasks by number may paradoxically imply increasing the practical importance of constraining choice, in order not to propagate variabilities that are difficult to include in equitable measurements. Candidates may be required to display comprehensive performance over the whole of an appropriately-defined range, rather than choose for themselves, preferred modes of performance from the restricted range supplied in examinations. Widely-based freedom of choice (in subject matter) and the requirement to cover a variety of task forms are indeed features of **IBO** *Internal Assessment* design in *Group 2 Languages,* though equivalence of candidate choice is problematic, as investigated and reported in the presentation, analysis and discussion of empirical findings in Chapter 6. In comparison with the freedom of choice of presentation for *Internal* Assessment, the inevitably-limited offerings of the *Written Production* design appear to accentuate the phenomenon[326].

In summary, key research issues investigated have related predominantly to questioning the design of 'authentic' tasks and the standardisation of such tasks in repeated formal assessments, of perceptions and understandings evidenced in candidates' responses, of the processes of successive moderations for attaining consensual commonality in assessor verdicts, and of issues for establishing validity, reliability and plausibility in the interpretation of data[327].

Grade Awards and the Relating of Scores to *General Grade Descriptors*

Gipps summarises conclusions for aggregating discrete component scores within any evaluation system, as demonstrating the incompatibility of the process with true criterion-referenced assessment. Justifiable evaluations emerging from aggregations are distorted by the phenomenon of 'compensation'. That is, weak performance in one discrete area of formal assessment may arithmetically be 'compensated' by strong performance elsewhere. This leads to the following, logical conclusion:

> "If strict criterion-referencing were translated into exam performances [.....], it would mean that the final subject level would be determined by the *worst* skill areas"[328].

As reported in Chapter 2, besides the complex arrangements published in the **IBO** *Vade Mecum*[329], *General Grade Descriptors* for grade-awarding, after moderation, serve as ultimate referents in evaluations[330]. For triangulation by **IBCA**, they are applied to each sample of candidate-production at given levels of performance, as an overall, criterion-referenced control of the effects of aggregation. This results in assessments of 'balance' for distributing weighted, componential scores across each tripartite, measurement domain, discretely graded as *Internal Assessment, Text-Handling* and *Written Production.* The *General Grade Descriptors* are designed to ensure criterion-referenced validation for the entire process of evaluation in all its outcomes. Final grades should therefore authoritatively and justifiably label all assessed language-productions. Hence at least theoretically, the **IBO** avoids distortions through 'compensations' in truly criterion-referenced, aggregated assessment, in accordance with Gipps and others' identifications. Discussion of the practical problems encountered in the course of research is provided in Chapter 6.

PART III

RESEARCH METHODS AND FINDINGS

CHAPTER FIVE

RATIONALES AND METHODS

Preface

In preliminary research, authenticity was broadly conceptualised as a referent for exploring issues of communication, assessment and evaluation in situated language use. Prior theoretical knowledge, framing re-interpretations of regularly-completed, classroom, homework and examination language, produced under **IBO** procedures and criteria for task-design and assessment, evolved under empirical investigation. Formal research questions arose from the results of pilot study, focussing perspectives and guiding the identification and selection of relevant evidence[331]. Data-collection and analysis could in this way, be comprehensive, yet coherently categorised. **IBO** criterion-referencing procedures remained unchanged, through multiple applications of moderation, to produce triangulated consensual evaluations. Without attempting to validate criterion-referencing as an assessment method in its own right, the number of variables requiring experimental control was thereby limited. The focus highlighted the conceptualisation and description of alternative criteria, designed to measure key features of authentic language use.

Devising reliable benchmarks for analysing situated production and exemplifying authentic expression required eclectic methods that consistently reference data to **IBO** documentation, candidate production and known theory. Three main routes were followed for investigating the validity of authentic, task-based language. First, samples of responses, produced according to **IBO** rubrics, were assessed by recommended procedure under conventional and experimental criteria, the latter specifically developed for research purposes. Observation and recording of **IBCA** assessment, moderation and evaluation practice formed the objectives of a second. The third required discourse analysis of the organisation's conceptualisations of authenticity, whether 'espoused' as theory for *Group 2 Languages,* or as theory in administrative 'practice'[332].

Overall, these complementary approaches to data-identification, creation, collection and analysis were simultaneously developed in scope and detail. As in *Action Research,* evidence was gathered and analysed in annual cycles, according to **IBO** examination schedules. The results led to progressive refinements of method.

The Scope of Empirical Research

By candidate registration numbers, French is a significant **IBO** *Group 2 Language*[333]. Empirical evidence was therefore primarily drawn from this domain[334]. The contemporary language, albeit 'Western' and Indo-European, is also significant for being discretely and explicitly defined, well-known and easy to reference to a little-contested 'standard'[335]. The choice adds clarity to the representation of language use through the elimination of inter-language variables and confinement to a single, though large, case study.

Within the range of **IBO** French programmes, performance data have been gathered, for practical reasons, solely from *Language B, Standard Level.* Ease of availability for procuring sufficient primary evidence and the time required for detailed familiarisation with such data restrict scope for a single project under a sole researcher[336].

Resource constraints additionally confine data-collection predominantly to recorded performances in speaking and writing[337]. Certain activities however, interlink these language-productions with listening and reading. The latter skills are unexamined either *per se*, or as elements for the psychometric measurement of comprehension. By design, the purest assessment of receptive knowledge and skill requires no display of language. Being non-communicative, it is extraneous to research purposes[338].

A priori considerations such as these structured and regularised data-selection. Confidentiality, in the context of accessing **IBCA** archives and analysing named-candidate performances, was respected without ethical or practical difficulty. With data readily available, personal identities were rendered anonymous, being otherwise irrelevant[339].

The predefined bounds of empirical investigation also determined the scope of literature reviewed[340]. Theories of testing and assessment informed content-selection and presentation for designing rubrics and tasks that stimulate and situate assessable performance. These frame formal, criterion-referenced administrative practices for evaluating validity of response. Guided by **IBO** programme philosophy, 'espoused' theory was contrasted with 'theory in practice', to allow critical appreciation of published examination and assessment schemes. The research methods therefore scrutinise the following: requirements and procedures for devising authentic tasks; task-standardisation for successive administrations of formal assessment; examiners' and candidates' perceptions and understandings of such tasks; and candidates' task-based responses. 'Familiarisation' with any given model and its consequential effects on validity, reliability and plausibility in data-interpretation were also investigated[341].

For practical reasons of scope, research in the latter areas, whilst relevant, was minor. Major analysis[342] was devoted to professional assessments and moderations of complete, **IBO**-selected sets of production from examining sessions for May 2001 and 2002[343]. Evidence for devising and standardising the 2001 examinations was also investigated[344], with **IBCA** procedures highlighting additional design constraints, linguistic and otherwise[345].

The Selection of Sources of Data[346]

IBO allocations of recorded performances were analysed as primary data. Formal assessments and experimental controls of these assessments, completed by the researcher, formed the greater part of the evidence. Independent, though methodologically-complex measurements of inter-rater reliability were largely excluded from investigation[347]. Data-collection by a single, **IBO**-trained and supervised rater unified the research design, though constraining potential for fuller validation through triangulation, and restricting the generation of reliable, plausible, wide-ranging and generalisable conclusions[348].

The disadvantages were partially offset by greater simplicity of plan. Evidence was triangulated through variation in vantage, with commonality attained through a single researcher's analysis of all data considered. In effect, the instruments were a single set of trained and experienced ears and eyes that shift in locus, but provide input for interpretation to a single mind, thus unifying the research[349].

For improving validity, reliability and plausibility in analysis, limitations in method were also partially offset by an element of inter-rater triangulation made possible through use of **IBO** administrative data[350].

For oral productions, **IBO**-moderated, teacher-based *Internal Assessments* provided comparative evidence. In some respects, such sources were inherently problematic, since performances were situated in differing, cultural, interlocutor-assessor and school-based contexts, either little, or uncontrolled by the **IBO**. In facilitating appropriate, interactive language use, the 'good faith' of teachers as *Internal Assessors* of their own students is only indirectly assessable and assessed.

For *Written Production*, triangulating perspective was achieved through sampling data from *Assistant Examiner Team Leaders* and *Chief Examiners*. In turn, these supplementary assessments had been derived by known procedure and criteria from larger samples of the researcher's own work as *Assistant Examiner*[351].

In exploring situated, authentic language use, 'typical' and 'anomalous' cases were identified in this way[352]. **IBO** moderations and evaluations were experimentally replicated, establishing unity of interpretative method as a referent for controlling validity and reliability. Other possible variables in approach were eliminated.

Common criteria and procedure allowed valid analysis of rater-reliability, not only from repeated assessment of candidate language-productions, both official and experimental, but also from comparisons of verdicts across discrete groupings of *Internal Assessment Moderators, Assistant Examiners* and *Team Leaders*. At each session, **IBCA** formally determines a co-efficient of reliability for each rater-employee, measured by organisational norms[353]. Qualitative representation, experimentation and reporting by the single researcher, were related to quantitative analysis and evaluation, with comparative calibrations established independently in **IBO** statistics. In this way, linkages between differing vantages became more evident, coherent, reliable and valid.

Material Excluded from Investigation

Pilot study conclusions recommended additional data-collection. Supplementary investigations deepened understanding, brought refinement to experimental designs, and are therefore briefly reported.

It was planned to survey student attitudes and approaches to task-choice in formal assessments. This would have provided data for response preparation and composition, either oral or written, and the checking and editing of outcomes in *Written Production*[354].

Further survey would have covered the attitudes and experiences of **IBO** task-designers and standardisers. It was evidently desirable to investigate processes by which understandings, consensus, commonality in standards, approaches to production and regularity in procedure might be established amongst groups of candidates and designers. As a result, the research could have been complemented with data-analysis of forms, content and rationales for the informal favouring of particular patterns of thinking about authentic language use.

Such investigation was left incomplete since, requiring additional resources, it implied extending boundaries and altering priorities. Conflations would have interlinked assessment with issues in pedagogy and learning, despite their relevance for teacher-practitioners and readers of the research. The focus on assessing language-production would therefore have been more diffuse.

Consequently, evidence was selected to permit as full a description, analysis and critique of programme-planning, administration and outcomes as possible, within the context of assessment and evaluation under a single, well-defined, **IBO** scheme. Included therefore are: the appropriate delineation of scope and boundaries for language and level; the investigation of rubrics, assessment-tasks and their standardisation; the use of language in criterion-referenced assessment; with the moderation and evaluation of results by published, numerical grading.

Unsought and unpublished **IBO** material for internal use was excluded. However, all regular research-reporting was copied to the **IBO**'s *Research Unit*, in order to maintain a productive relationship with the primary-data provider, and in fulfilment of initial agreements securing the organisation's willing co-operation[355]. On reception, further archive material was made available as potentially relevant to the research, a fact suggesting repeated, if tacit approval of methods, investigations and interim results. Significant omissions therefore appeared unlikely.

The Description and Experimental Analysis of Data

Besides scrutiny of documentation identified in Chapter 2, complementary, empirical approaches were also used, one for each discrete data-collection exercise[356]. The first sought to establish valid, reliable benchmarks for a range of experimental assessments, allowing comparative analysis of criterion-referenced evidence. In graduated sets, qualitative criterion-descriptions were quantified by number, ranging in polarity from maxima for highly-substantiated, incontestable evidence of authentic expression, to minima for its complete absence.

An instrument was devised for identifying and describing key features of authentic expression, facilitating categorisation and analysis of exemplary assessments, whether oral or written[357]. Experiments required analysis of recorded language-production from the research data-base. For each examining session, complete sets were processed by the researcher as assessor, exercising simultaneous functions as **IBO** *Internal Assessment Moderator* and *Assistant Examiner.*

Through generating supplementary data from a common body of evidence, comparable analyses could be triangulated. **IBO** assessment and evaluation paradigms were thereby opened to greater critical purview.

In simulating **IBO** philosophy and practice, the varied triangulations of experimental research enhanced the validity and reliability of the interpretative processes they entailed. Unchanging assessment criteria and methods were consistently compared, with comparisons founded on common and stable corpora of evidence, interpreted through applying common procedure. Primary data were collected from similar assessment contexts on different occasions, from different sets of candidates, in different combinations, according to **IBO** allocations. Validity was partially controllable and controlled through repeated, longitudinal applications of identical triangulating method, over three and more similar, formal examining sessions[358], as in **IBO** practice, where successive moderations, albeit by additional raters, build consensual verdicts through repeating assessment processes in their entirety[359].

The use of consensual, criterion-referencing method is apposite in its 'authenticity' as the explicit, attempted simulation of likely processes in any communication. Assessors as alternative listeners and readers evaluate communicative qualities for extended and elaborated productions of language, albeit in non-interactive fashion under traditional forms of assessment. Hence, experimental triangulation as multiplications of perspectives on single pieces of production was deemed appropriate for exploring assessment schemes that value authentic expression. Through altering the perspectives of assessment, experiments created greater relief in understanding, heightening awareness of central issues, whilst remaining grounded in theories of authentically-based, communicative language use.

The Measurement of Authentic Language Use

The research instrument employed original categorisations of authenticity, derived from the conceptualisations of Van Lier (1996)[360]. For this author, communication successfully realised via the common language of two or more interlocutors as speakers and listeners, or two or more partners as writers and readers, displays evidence for authentic expression, distinguishable in ten discrete, yet interlinking perspectives. In themselves, these categories have no pre-determined, quantitative value, the model being theoretical and unconcerned with issues of assessment. Indeed, whilst individual components may be demonstrated as qualitatively valid, conflation through aggregation may be problematic. The overall validity of 'weighted', totalised component scores is irrelevant to Van Lier's purpose and unaddressed. For initial theoretical and experimental purposes therefore, conceptualisations were developed instrumentally, and listed in three groupings according to the author's specifications. In summary, but without prioritisation in listing, these were classified and explained as:

Evidence for 'Curricular Authentication'

- *Creator authenticity as notions and linguistic realisations of 'self',* focussing attention on the personal and unique 'voice' of each producer of language.
- *Creator authenticity, as perceptions of 'other' as interlocutor, audience or reader,* illustrated by attempts to motivate participation in communicative interchange through personal strategies or discrete tactics that retain listeners' or readers' attention.
- *Finder authenticity, or the resourcefulness of communicators in finding material for communication,* giving evidence for the development of recognisable agency in the selection and manipulation of specific objects of awareness, sourced in worlds outside 'self'.
- *User authenticity, or recognitions of 'other' as listener or reader, and as focussing attention and linguistic interaction through respect for commonly-acquired social traditions and communicative convention, thereby allowing coherent initiations and continuations of communication*: there is evidence of purposive response to set stimuli and to prompts sourced in the initiatives of 'other'.

Evidence for 'Pragmatic Authentication'

- *Authenticity of context, or the willingness by partners in communication to share culturally-situated perspectives through initiating communication in recognition of, and respect for the traditions and conventions of collaboratively-modifiable culture:* evident are agreements, explicit or implicit, interactively to share communication and so construct extensive and extendable social relations through language. There is no suggestion of 'self'-determined, one-sided closure of communication.
- *Authenticity of purpose, or transparency and self-awareness in choices of expressive genre, and communicative message:* identifiable are intentional facilitations of changes in perspective and knowledge amongst audiences of the text created, and on occasion, reflexively in 'self'.
- *Authenticity of interaction between partners in communication, or recognitions of power in questions of balance, 'convincingness' and validity, determining communicative quality in social relationships between speakers and listeners, writers and readers:* evident are accommodations of 'self' to 'other' in processes of continuous change, and recognitions within 'self' and in 'self as other', of ability to guide this development.

Evidence for 'Personal Authentication'

- *Existential authenticity, or social constructions and expressions of 'self' through (communicative) actions,* focussing attention on awareness of the uniqueness of personal 'voice', or negatively, on avoiding overt plagiarism through its absence as evidence.

- *Intrinsic authenticity, or recognitions of self-determination as significant process in revealing continuous operations of socio-temporally situated choice:* attention is focussed on evidence for active, metacognitively-conscious selections in language-productions.
- *Autotelic authenticity (after Csikszentmihalyi[361]), or experiencing and expressing 'flow' as 'optimal experience', relating linguistic coherence and psychological balance to the inner mental worlds of subjects:* attention is focussed on evidence for committed concentrations of awareness on jobs at hand, with intentional strivings through communication to satisfy personally-chosen goals, without intrusive distractions, or irrelevance.

The Design of the Research Instrument

To facilitate data-categorisation, Van Lier's concepts were grouped into sets of graduated descriptors, simulating assessment within the **IBO**'s qualitative, criterion-referencing, interpretative tradition. Hence performance at one of five levels of qualification gave rise to approximate quantifications in similar number, evaluated through comparable procedure[362]. However, to reduce scope for variation in judgement, no leeway within each level was provided for further evaluation by subjective 'adjustments' of one point, as in the **IBO**'s referent model[363]. In typifying attainment, the range progressed discretely from a minimum stipulation of evidence as "*no[ne.....]*", to a maximum that is "*abundant*", through three intermediate levels designated respectively as "*little*", "*adequate*" and "*significant*", and illustrated in **Appendix 3**. Through preserving Van Lier's ten categories, with further discrimination for each category described in five discrete, single-point levels of performance, a simple, though unrationalised and unweighted quantification of attainment was equated to a maximum possible score of fifty points. High levels of validity were retained, and reliability enhanced through the reduction of scope for subjective variance in classifying interpretations by value. Language was assessed by appropriate matching to discrete, single-point levels of description, with no variation of attributed quality possible within each level.

After experimental assessment under this form of calibration, the design was modified to improve consistency in creating valid and reliable triangulations. Hence distinctions between descriptions were sharpened, yet left ‘idealised’ as qualitative categorisations of hypothetically ‘typical’ language-productions. Experimental validity was thus enhanced, even though values for each measure remained dependent on the reliability of the researcher as assessor in interpreting criteria and matching criterion-descriptions to the evidence of productions[364].

With memory-retention by listeners and assessors influencing the assessment of communicative quality in real-time oral interchanges of up to fifteen minutes’ duration, and with written texts of a recommended minimum length of 250 words for example, there was little data found clearly to distinguish descriptive quantifications of “*significant*” and “*abundant evidence*”. The major components of each criterion were therefore reduced to three by conflating descriptions of ‘*significance*’ and ‘*abundance*’. Two complementary levels were added to identify extremes at either end of the scale, one negative, signalling a complete absence of evidence, and the other positive, for interpretatively-incontestable displays of competence. In this respect, rater-reliability was improved by requiring judgements to relate not to one of five, but to one of three, more clearly-differentiated, single-point categories, applicable to the description of the large majority of cases. The inherent subjectivity of rater-interpretation was more constrained (and thus less variable), with numerical criterion-scores more precisely matched in value. The refinement also provided indicators demonstrating conditions of 'inauthenticity'. An excess of extreme scores could reveal inappropriate programme and level selection by candidates, with tasks being either too 'easy' or too 'difficult', a factor imperilling the viability of appropriately-contextualised, authentic communication and the validity of differentiated programme and level-based, criterion-referenced assessment.

Through improving single-rater reliability, it was anticipated that inter-rater reliability could also be enhanced, though as stated, such research was not completed. Using the refined instrument, assessments may be summarised as judgements for which there is an *unsatisfactory*, *satisfactory*, or *more than satisfactory* provision of evidence, within each criterion. Thus over ten criteria, approximate quantifications of attainment equate to a maximum aggregation of thirty points. Compensatory bonuses and penalisations, where clearly evident, permitted 'adjustments' for fine-tuning results, more precisely discriminating different performances. Score-totalisations allowed direct, though crude comparison with the maxima of thirty of the referent IBO model[365].

In this way, the more refined experimental instrument required graduations recording the supply of evidence as "*little*", "*adequate*", or "*abundant*" in cases that were not extreme. Reliability, limited by subjective variability in assessor interpretations, was improved insofar as essentially quantitative categorisations were clearly distinguished in tripartite fashion, and made applicable to most empirically-researched examples of production. Van Lier's ten features of authentic language use were retained, enhancing through such refinement, the stability, justification and validity of simply-quantified assessment verdicts, even though each was evaluated solely in whole-point scores.

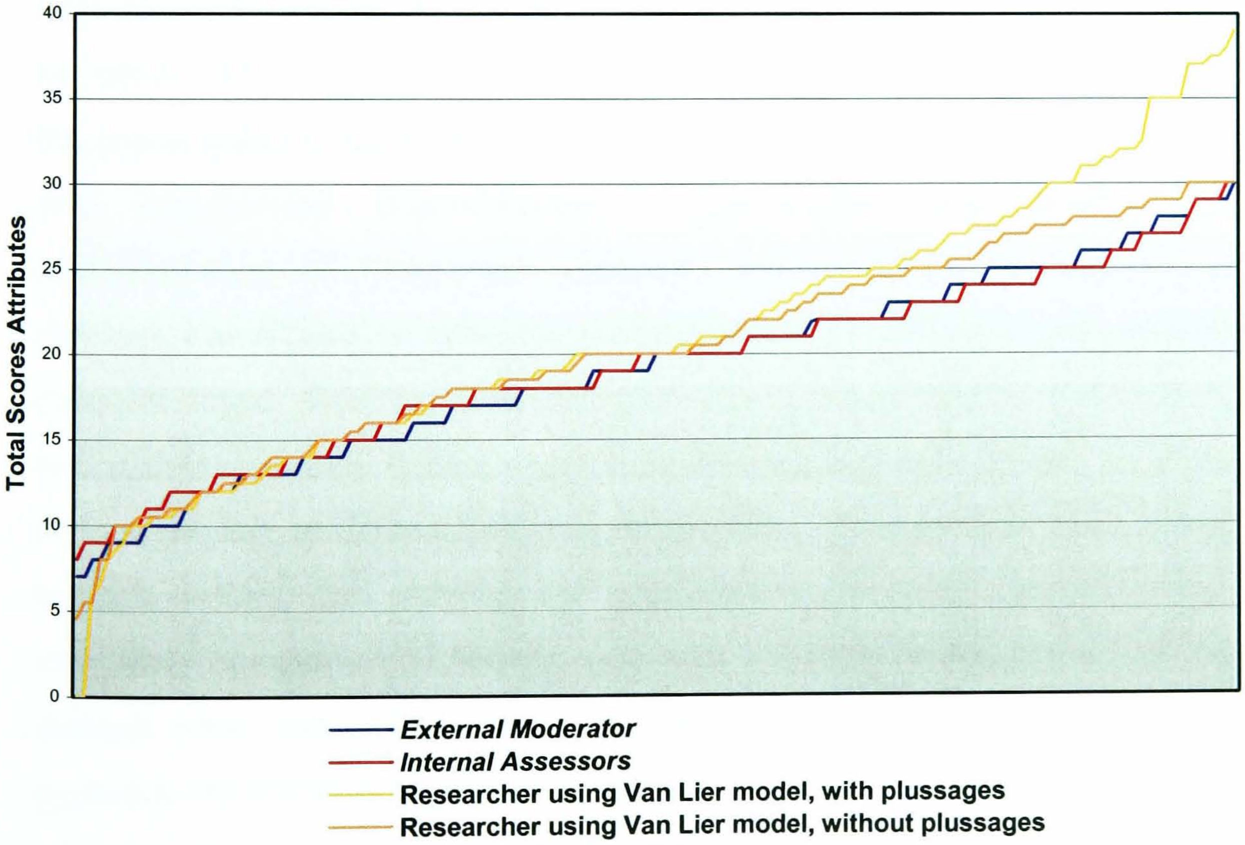
Hence, qualifications designated as "*little evidence supplied*", were valued at one point; "*adequate evidence supplied*" was quantified as two; and "*abundant evidence supplied*" scored three. Aggregating all scores to a possible maximum of thirty for each production, reduced the need arithmetically to manipulate calibrations for direct comparison with the **IBO** scheme. Potential sources of rater-error and of increased distortion (despite the absence of rationales for fine-tuning the weighting of aggregated totalisations) were reduced in number. Straightforward and illuminating comparisons between evidence analysed and triangulated under existing and experimental systems were easily, if somewhat crudely made for the purposes of exploration and illustration.

Indeed, as shown in Chapter 6, the bulk of analysed data provided unproblematic examples of attainment at one of the three major levels described. In practice, few 'adjustments' to total scores through applying bonuses (or penalisations), as categorised for extreme cases by the refined model, were necessary[366].

Representing a total of 150 *Internal Assessment* candidates, sampled for the May 2001, 2002 and 2003 evaluation sessions, the major data was graphically triangulated to produce four-way comparisons, enhancing understandings of validity and reliability in assessment-criterion design and application, more than is the case for *Written Production*. Through generic comparison of alternative systems, the overall range of scores awarded by teachers as *Internal Moderators,* or by the researcher as *External Moderator* for internally-assessed, oral language-production, showed little significant deviation[367]. Each level of attainment indicated by a distribution line (whether 'low', 'average' or 'high') appeared directly comparable and stable, as displayed in *Figure 5.1.* Indeed, through allocating 'plussages' of one point for exceptional provision of evidence satisfying any given criterion, or subtracting one point for failure to provide such evidence, (allowing a total of forty points), improved discrimination of performance extremes enhanced differentiation across the range, and notably in the upper half recorded.

Figure 5.1

Total scores awarded by *Internal Assessors* and the *External Moderator*, (sessions for May 2001, 2002 and 2003 aggregated), against assessment derived from Van Lier, by the researcher.
(Sample size = 150)



In such graphical representation, ideal results for typifying schemes that effectively differentiate the values of unique productions should describe diagonal, straight lines. Here, scores derived from research data were plotted in order of increase, to show overall variation across the differing systems. They were measured against the constants of **IBO**-selected groups of teacher-assessors, the single researcher as *Moderator* and *Examiner*, and stable samples of candidate productions, recording interactions through authentic expression according to common, **IBO**-defined rubrics and an **IBO**-selected, overall range of tasks[368].

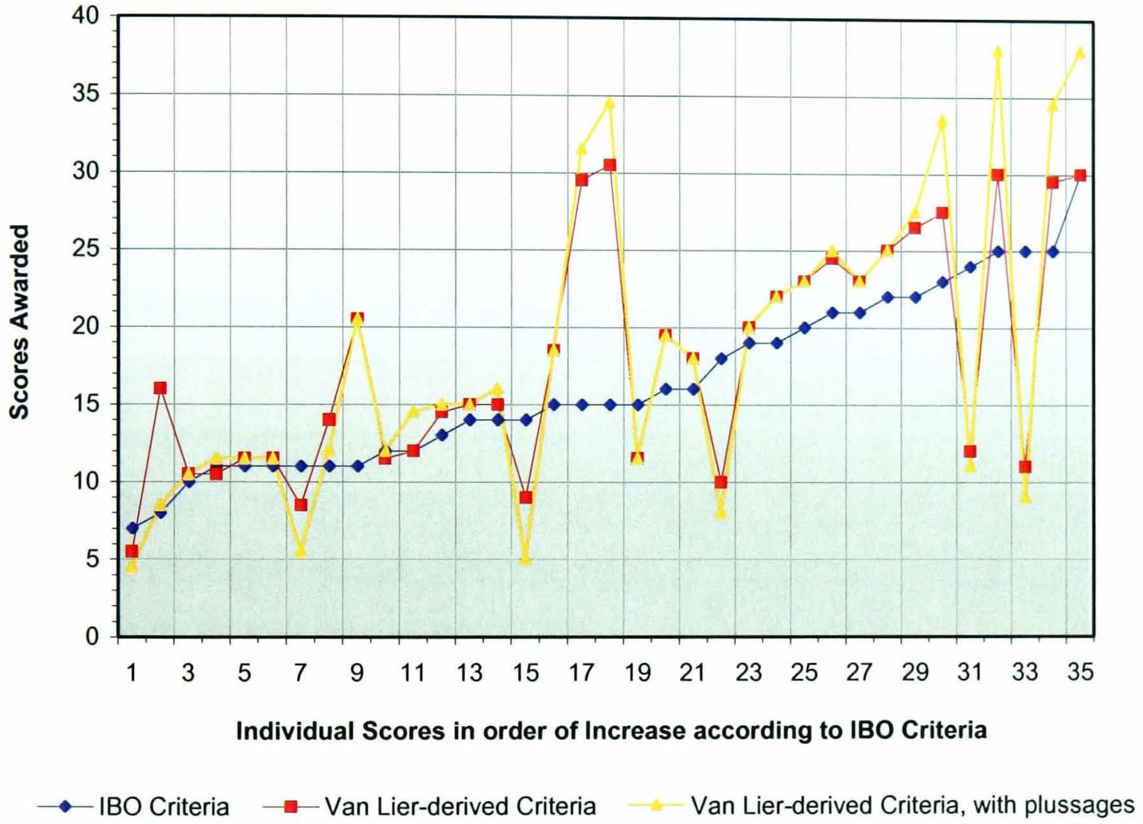
Graduated measures of authentic language use may be illustrated thus, albeit with artificial constraints. Nonetheless, in interpreting graphical representations, it should be recalled that truly authentic performances are irreproducible over time and context, being individualised within particular socio-cultural and temporal situations by 'self' in interaction with alternate 'selves' as 'other'. They will always vary, even if variation be small. Hence, the more diagonal and straight the lines described, the more systems approach ideals for evaluating authentic performance with appropriate discrimination. Conversely, the greater the representation of horizontal *plateaux*, the more the production of different candidates in different performances on different tasks and at different times, albeit for assessment under identical rubrics and criteria, and with assumedly stable assessors determining appropriate verdicts, in fact remain *undifferentiated* in evaluation. It is evident that in any system reliant for validity on aggregating discrete, appropriately-weighted, componential scores, any two, identical totalisations may be derived from widely-divergent, individual component scores. In such systems, discrimination is weak, since equally-valued performances for any given task may represent significantly different qualities of performance.

Through this method of representation, scores derived from **IBO** criteria by *Internal Assessors* or the *Internal Assessment Moderator* could be closely correlated. This suggests a high degree of reliability between group and researcher in relation to quantitative evaluations, though tendencies to higher scoring in the upper range and lower scoring in the lower range were evident for the researcher, producing a slope that is nearer to the ideal[369]. The trend is graphically indicated by the slight bias to the left or right of the blue line in comparison with the red. Comparing applications of **IBO** criteria with experimental data also demonstrates close correlations, though experimental assessments are scored slightly higher overall (in researcher applications, at least). Further exaggeration was produced through applying plussages, more precisely differentiating quantifications of descriptions, especially in the upper range (as expected in totalising to a maximum of forty points), but less so in the lower where scores remained close. Furthermore, the overall diagonal described in this 'enhanced' range for quantifying large sets of individualised productions, better approached the ideal of perfect discrimination[370].

With comparative graphical representations displaying results allocated to individual candidates through different assessment schemes, clear identifications of aberrance were to be expected. From these, examples of performance were isolated for detailed description, analysis and discussion. Indeed, during the processes of **IBO** assessment and evaluation, many examples were revealed in the research archive. In contrast to *Figure 5.1* where all task responses, being individual, were necessarily aggregated, a typical sample for a single, candidate-chosen task, illustrating scores in *Written Production,* is shown in *Figure 5.2*[371]. Anomalous cases appear as outliers from the diagonal, with the extended model (including 'plussages') more clearly emphasising such anomalies, especially in the higher ranges of scores derived from **IBO** criteria[372].

Figure 5.2

Comparison of Scores Awarded in a single *Written Production* task
(Sample Size = 35)



The Research Instrument in Use

From experimental usage, Van Lier's componential classifications of authenticity proved meaningful, discriminatory, broadly in accord with the referent model, and capable of indicating clear anomalies. Developed for triangulation with **IBO** criterion-referenced assessments, the instrument distinguished language-performances in coherent graduations, quantifying qualities of authentic expression through a system commonly applicable to written and oral production.

However, in denoting holistic vantage points for multi-dimensional viewings of individual pieces of evidence, conceptual categorisations seem in certain cases, partially to overlap. Aspects of *User Authenticity*[373] for example, emphasising recognition of, and respect for culture common to language producers and receivers, are similar to those for *Authenticity of Context*, as "the willingness by partners in communication to share culturally-situated perspectives".

In distinction, *User Authenticity* may be applied to culturally-situated, content-rich, task-based responses for specifically-chosen scenarios, with *Authenticity of Context* focussing assessment attention upon initiations of communication in recognition of, and respect for linguistic tradition and cultural convention, all-enveloping in setting yet collaboratively-modifiable in essence. With French as the medium of communication, this is no micro-culture defined by a particular genre or task, as in *User Authenticity*, but the world of francophones, sharing the use of French as a common language.

Similarly, *"Evidence for 'Personal Authentication'"* emphasises operations of choice, committing focused attention and effort on choices once made. These features are predominantly psychological and sociological in dimension, with assessors seeking markers for *'Existential', 'Intrinsic'* and *'Autotelic Authenticity'*, and redirecting vantage in holistic fashion from the more language-form and content-based criteria for *Evidence for 'Curricular'* and *'Pragmatic Authentication'*.

Notwithstanding the blurring of boundaries between discrete perspectives, experimental data provided sets of evaluations from unified production-domains, with analyses allowing comparison of assessment systems and rater-judgements. Communicatively-interactive relations between speaker and listener, or writer and reader, were measured for authenticity with a common instrument. It was therefore possible to describe and quantify evaluations for any task-based, culturally-situated communication, by measuring aspects of the relationships between language 'producers' and 'receivers', even though validity and reliability in quantifying transformations by numeric value remained imprecise or problematic[374]. The essentially subjective, interpretative processes establishing relations between 'self' and 'other', eschewing itemised, positivistic and psychometric assessments of linguistic data, were measured under triangulating and empirically-based, holisitic perspectives, regardless of particularity or level of task, or means of expression.

Enhancing understanding through triangulation exercises such as these, did not in itself, lead to the production of valid and reliable data as substitutes for **IBO** assessments. Indeed, in practical terms, the experimental instrument remains rudimentary and capable of further refinement, being dependent in use on assessor interpretations and unproblematised linkages matching qualitative observations to numeric scorings. For quantitative evaluation, investigation of validity either in weighting by categorisation, or through positivistic norm-referencing[375] had been excluded from research. However, overlappings of criterion-descriptions neither jeopardise the comparative validity of experiments designed to *explore* authenticity in assessment, nor exclude coherent critique of an existing system, since no development of alternatives is posited. The model may be insufficiently grounded in wider theories of validity for aggregating discrete component marks in totalising final scores. The experimental instrument is used with caution in this respect. Through attributing scores to criterion-relationships, even though weighted by assumed values that have neither been theorised, nor empirically researched, its invalidity nevertheless, is constant. Given awareness of limitations in approach, developing insight remains unencumbered[376]. Methodological consistency in respecting official process and measuring full sets of assessments longitudinally over time, favoured stability of interpretation. The experiments created stereoscopic views for investigating **IBO** procedures and assessments within an intrinsically valid framework[377].

Use of a single individual for creating the empirical data-base limited variability in assessor-perspective. With the number of archived examples of candidate language-production increasing over time, inherent problems of interpretative validity and interrater-reliability were progressively reduced. Such reduction occurred through saturating the theorisation of authenticity with analyses of primary evidence. Deleting, modifying, further refining and indeed adding new categories were procedures thereby ever more firmly grounded in processing samples of performance, sourced amongst **IBO**-selected, though effectively randomised centres and their candidates, engaged in 'real' sessions of *Diploma Programme* assessments and examinations, and assessed by the researcher as *Assistant Examiner*[378].

Linking Assessment to Performance-based Task Completion

At this stage, the taxonomy of experimental criteria developed for measuring evidence of authentic language use requires re-consideration of assessment theory, and in particular, the literature devoted to issues in criterion-based assessment. As Bruner (1998) has demonstrated, the agentive mind seeks out collaboration in culturally-embedded, problem-solving projects, established with other minds through communicative performance. Such collaborative acquisition and use of skill is primarily linguistic in nature, requiring discursively-based dialogue, uniting language production with reception. Practice is performance in ways that are never discrete, since it combines the use of at least two fundamental skills: one receptive and one productive. Indeed, skilled, culturally-embedded, albeit agentively-produced discourse between listeners and speakers, writers and readers, forms the prime object of research assessment and evaluation. In particular, as has been seen, determining relative weighting-values for discrete assessment categories in ultimate aggregations of scores, remains an unexamined measurement problem. Without such values, justifiable distributions cannot be established across the four domains of language use. In the following recountings of method, mixed-skill performance assessments, rather than final evaluation therefore form the focus of attention.

Assessing Reading and Writing

Approximately 150 examples of written language were analysed per examining session[379]. Over the course of the formal research, more than 450 scripts were assessed under duties as **IBO** *Examiner* and *Internal Assessment Moderator*, of which 60 were formally correlated with the judgements of other **IBO** assessors, through the moderation process[380]. A total of approximately 1,050 scripts were assessed in the period 1996 – 2003.

For *Written Production*, analysis followed **IBO** practice[381]. This required assessment and rating according to procedures and criteria, described in Chapter 2[382]. For experimental purposes, most individual scripts were immediately re-assessed, replicating standard procedure though applying Van Lier-derived criteria. In some cases, given constraint in working to invariable, **IBCA** schedules, the process of re-assessment was completed neither consecutively nor contemporaneously. On occasion, certain examples were reconsidered through further replication of procedure, after a lapse of a year, or more. However, all empirical evidence presented was assessed, re-assessed and evaluated within the period set by the three-year bounds of the project.

In supplementing meanings for enriching triangulations with this data, practical constraints, undesirable as a source of methodological inconsistency, were nevertheless insufficiently significant to threaten validity, or alter the general aim of developing insight into the feasibility of measuring authentic language use. Creating longitudinal, temporal dimensions in experimental assessments advantageously confirmed the validity and reliability of the single researcher's judgements[383].

Procedural arrangements for determining validity in preliminary conclusions and examiner/moderator reliability per assessment session, meant that between one and six examples of candidate production were processed in any uninterrupted, **IBO** application, with a maximum of twenty in any single day, throughout the twenty-eight day period officially allocated for these purposes. Following assessment of between approximately 60 and 80 scripts per examination, twenty were sampled for moderation according to **IBO** requirements, including examples from as wide a range of scores and centres registering candidates as possible[384].

Following each annual collection of data, copies of the research instrument were dispatched for comment to certain **IBO** employees as interested colleagues[385]. On reply, the experimental criteria were further refined for re-use in subsequent assessment and evaluation exercises.

Assessing Listening and Speaking

For additional perspective on **IBO** conceptualisations of authenticity in task-design, and response-assessment, two complementary sources of evidence were investigated, extending the original focus of the research as a study of written production.

The method used in processing the former involved triangulating assessments of authenticity through comparison of evaluations for written productions with those for moderated *Internal Assessments*. This contrasted language skills and knowledge in time-dependent interactions in listening and speaking with relatively time-independent activities in reading and writing. Different **IBO** strategies for designing criteria and recommending assessment procedures in the differing language domains could be compared. Supplementary data-collection allowed description and analysis of presentations proposed by candidates themselves, (albeit according to broad guidelines and specific rubrics set by the **IBO**), in which attention focussed on interactive listening and speaking as initiative and response in extended discourse, and on assessment, moderation and evaluation for this oral production.

As **IBO** *External Moderator* for *Internal Assessment*, the following were assessed and moderated: 55 samples of oral production by *Internal Assessors* and candidates from Canada, The Netherlands and the United States in May 2001; and 51 similar samples, solely from the United States, in May 2002[386].

In these assessments, previous data-analysis methods were replicated. That is, candidate work was analysed according to **IBO** criteria and procedure, with experimental re-assessment immediately following[387]. The process produced comparable, quantitative data derived from at least four, distinct listenings, with grades allocated by each internal teacher-assessor, scores determined by an *Internal Assessment Moderator*, sampled re-assessments by **IBCA** *Team Leaders* and experimental analysis. The differing, quantified evaluations were recorded in varying forms of graphical representation[388], as displayed and discussed in Chapter 6.

It should be recalled that recourse to a single assessor reduced measurement variability, with procedural reliability thereby increased. This permitted short, mid and long-term investigation of degrees of longitudinal stability in a single rater's comprehension, applications of procedure and resultant interpretations[389]. Even though for concluding, validity and reliability may be enhanced through triangulation by experimental replication, with data derived from trained *Examiners* and independent *Moderators* employed for this purpose, thus multiplying the number of assessors as readers and audiences for performances, such strategy would considerably have increased the complexity of the research[390]. Within constraints of time, finance and defined scope for the project, this extension was deemed unfeasible.

Nonetheless, use of *Moderator* and *Assistant Examiner* reliability statistics, partially mitigate the limitations of recourse to a single source for much data-production. Co-efficients are drawn up for each **IBCA** employee at each examining session. For the researcher, longitudinally over more than three years of examinations, they measured and recorded performance as acceptably stable and 'typical', according to organisational standards[391]. Hence one externally-validated measure of consistency signifies a major point of reference. It is thus claimed that experiments permitted development of greater analytical depth, despite certain methodological constraints. Valid insights for inferential conclusions were adequately reliable, given varied triangulations comparing evaluations of increasing numbers of stably-recorded language-productions, replicated over time.

This research method was reapplied in the annual processing of approximately 50 examples of oral discourse, within two successive, examining-sessions (May 2001 and 2002)[392]. In due course, a total of 107 examples were analysed. However, in contrast with procedures adopted for *Written Production*, official and experimental evaluations were completed uninterruptedly over the **IBO**'s twenty-five day schedule. For maintaining validity and reliability of moderator performance over such short periods, between one and eight samples of oral productions were analysed in any single day.

In further contrast, samples of eight were selected according to stipulations on completion of duties per session, ('problem' cases posed by the failure of candidates and *Internal Assessors* to respect assessment, being excepted and excluded, under official procedure)[393]. Such samples similarly reflected as wide a range of candidate-scores and of registered centres as possible, with re-moderation by supervising **IBCA** *Examiners*, replicating measurements for validity and reliability, as described for *Written Production*.

Further Developments of Method

To summarise, in applying assessment and evaluation criteria, common methods and procedures were adopted, first under **IBO** requirements, and then in experimental re-assessments, according to a further strategy. In oral presentations and discussions, as well as written productions, individual texts were first considered in their entirety. Thus as 'authentic' listener or reader, situated language could be informally appreciated according to the propositions advanced orally, or the tasks selected for writing. The technical requirements of listeners and readers as assessors or moderators were initially ignored and replaced by those of an interested party as an individual, representing unspecified audiences or readerships. Further listenings or readings followed, occasionally with necessary 'inauthenticity', in the sense that reasons for repetition on the part of 'other' are external to acts of listening or reading in themselves. Rationales derive most often from the requirements of assessment, in these cases from sources outside the 'self' of situated listeners or readers. The positioning is defined by specific **IBO** assessor-roles, with the organisation itself further positioned in supervisory roles as the authoritative and final arbiter of all evaluations. Thus, repeated assessments of listenings or readings were constantly referred to the tabulated, **IBO** criteria, and to those of the experimental model[394].

Assessment completions were also measured and recorded in time. This was found to vary between a minimum of twenty minutes (for a large majority of straightforward cases) and a maximum of thirty minutes. Only exceptions required more, mostly in written productions where handwriting-styles were difficult to decipher.

With a fifteen-minute duration stipulated for oral production, the process extended to approximately thirty minutes, including two full listenings and the appropriate, simultaneous processing of assessment information. Problematic cases usually concerned difficulties with technical qualities in recording, poor candidate audibility, or disrespect of set requirements by *Internal Assessors* dominating productions in quantity of recorded speech, at the expense of candidates.

The results at each examining session were sampled and dispatched to further **IBCA** *Moderators*, as previously described[395]. *Subject Reports*, and on request from named centres, *Individual School Reports* for full sets of candidate-performances, were composed and considered by *Chief Examiners* at moderation meetings, under **IBO** regulations.

Certain copies, if difficult to interpret, were therefore processed after a third, or further listenings or readings. Any unresolved cases were dispatched to co-ordinating *Moderators* or *Examiners* for re-assessment as "problematic", under standard procedure[396].

However, given smaller samples for *Internal Assessment,* research was completed concurrently, both under **IBO** and experimental schemes.

Supplementary data relating different modes of language-assessment to differing, oral and written components of a given programme under common criteria and procedure, were thus integrated within the research as a whole. Samples were collected simultaneously, longitudinally over time and 'latitudinally' in an increasing body of evidence for triangulating rater understandings of validity and reliability in assessments.

Interpreting the Experimental Criteria

In experiment, application of the Van Lier-derived criteria required the assessor to identify linguistic evidence according to choice of task and genre for response, interpretatively typified in summary as follows:

- *Creator authenticity as notions and linguistic realisations of 'self':* assessment requires evidence of linguistically-successful particularisations of identity. An individual as producer of language must be heard and revealed, for example in the recounting of autobiographical incident, personal attitudes, emotions, dilemmas, expectations for the future, amongst others, thus allowing the construction of originality and avoiding plagiarism.
- *Creator authenticity, as perceptions of 'other' as interlocutor, audience or reader:* assessment rates the attempt to maintain participation in communicative interchanges with appropriate, linguistic devices. There will be for example, evidence of personal strategies or discrete tactics for retaining and developing listeners' or readers' attention through the use of appropriate content, appropriately adapted by form, leading to 'satisfaction' overall, the text establishing relevance to its audience.
- *Finder authenticity,* where evidence is displayed in the content of responses of selection and manipulation of sources of knowledge appropriate to the chosen task and its genre, in intellectual, cultural, or emotional terms, and so forth, and sourced from worlds outside 'self'.
- *User authenticity,* where linguistic evidence of purposive adaptation of content as appropriate in response to chosen tasks

within their defined socio-temporal and cultural situation is assessed.

- *Authenticity of context,* or evident, interactive construction through linguistic means of extensive and extendable social relations. Text-production is acknowledged as response to initiatives from 'other' with precise reference to tasks and situations chosen. There is linguistically-indicated awareness of effects on response-receivers, as markers of interaction. There will be no evidence of sustained irrelevance, evasiveness (unless appropriate) or attempts to close down channels of communication.
- *Authenticity of purpose,* or structural organisations of content with linguistic evidence for promotions and facilitations of changes in perspective and knowledge amongst audiences of the text created, and on occasion reflexively in 'self', thus granting the production a quality of 'convincingness'. The receiver knows easily, why it has been produced.
- *Authenticity of interaction,* or linguistically evident accommodations of 'self' to 'other' in processes of continuous change, marked through the unfolding of the text. There is assessable evidence of responsive recognitions within 'self' and in 'self as other', of ability and will to guide this development.

The remaining criteria under *Evidence for Personal Authentication* are more psychological in focus, as has been explained, and therefore more purely dependent upon the experientially-based, subjective and personal interpretations of overall 'effect' in communication, developed by listeners or readers. In all successfully-realised, authentic acts of communication, as claimed, aspects of particularised intersubjectivity will be difficult, if not impossible to generalise as predictive statements typifying empirical evidence. Retrospectively however, *after* reception of any given text, these are often both clear, and easy to define amongst the myriad of possible particularities.

Observation and Recording of *Grade Award Meetings*

These activities formed a second strategy for data-collection. The proceedings of **IBCA** moderation and evaluation meetings were recorded for *French, Language B* in December 2000, and *German, Language B* in June 2001[397].

The researcher observed procedures according to **IBO** requirements. Unofficially, this included minute-taking. He was also invited occasionally to participate with comments on procedure or on particular evaluations, as if in moderation[398]. Meetings were therefore recorded by the researcher acting as a secretary, producing **IBO** reports also useful to the research[399]. Given time-limitations for attendance, semi-structured interviews with participants for verification and supplementary detail took place informally, at intervals between moderation sessions. As a result, they were only recorded in notes from memory, subsequent to each session[400]. Such unplanned constraints proved insignificant, with observation intended to record previously undocumented procedure. Indeed, with reports copied to the **IBO** for further verification, feedback was invited and occasionally supplied by participants[401]. Thus, a more tightly-focused reiteration of observation was developed for the same programme at the same levels, albeit in differing, yet complementary domains. For establishing commonality of process as suggested by the **IBO**, moderation and evaluation for *German Language B* were likewise observed at the *Grade Award Meeting* of June 2001[402].

Observations were deemed to have met objectives, as recounted in Chapter 6. Data-collection was substantial and enhanced by the exclusion of scheduling for pre-identifying items of particular relevance. **IBO** approval of researcher-reports, submitted in the role of *Teacher-Observer*, partly endorsed conclusions that records were consistent, accurate and comprehensive; in this sense, they were 'typical'. During observation, manuscript notes had regularly been offered to participants for cross-checking. On further request, they were provided for perusal and supplementary comment. They were also presented to the **IBO** *Director of Assessment* at the conclusion of each session[403]. No problems were encountered[404].

Research Data and the Design and Standardisation of Tasks

Data was also collected for describing and analysing **IBO** documents, made available after initiating research. Once trust in matters of security and confidentiality, had developed in relationships between researcher and **IBCA**, the latter suggested and facilitated access to archive material normally unobtainable, being for internal administration and 'sensitive' to the design and standardisation of 'high stakes', public examinations. This access allowed identification of further issues of authenticity in examination contexts, with examples researched for the May 2001 session of *French Language B*. It also highlighted general concerns in task-design and standardisation *per se,* both longitudinally across time within the French programme, and 'latitudinally' across a range of languages assessed at common levels[405].

The method adopted for researching such unpublished dossiers was to examine the archives for items informing the research questions. Particular terminological usage was noted in proposals of scenarios for 'authentic' tasks and promotions of authentic expression, with verbatim transcription of examples of special interest, recorded for subsequent reflection, analysis and discussion. In all relevant cases, the method raises issues of selection, interpretation, dissemination and referencing convention, together with considerations of authorship and intended audience, though these lie outside the project's bounds. Given the evidence, such issues were neither significant, nor relevant to central purposes. For ethical reasons however, as well as in satisfying **IBO** requirements, all notes were shown to the *Director of Assessment* and copied to the *Subject Area Manager* for *Group 2 Languages* prior to leaving **IBCA** premises, after each research session. In this context, no concerns were raised. Indeed, the absence of comment may reveal tacit approbation that data-collection was accurate and unbiased, with no significant distortion in representing organisational activity in this area.

From the evidence collected, requirements and constraints relating to **IBO** conceptualisations of authenticity could be identified and described in detail. Critical examination of this supplementary documentation served to refine grounded understandings developed through alternative method in the course of research, as recounted. In the absence of contrary evidence, it was assumed that analysis of programme-design practice and implementation requirements would prove resistant to selective distortion or inappropriate interpretation, once compared with assessment outcomes for texts that realise authentically-situated language use[406].

However, subsequent to observation and with assessor experience of moderating processes, this aspect of method particularly focused research attention on the review, with close, critical analysis, of issues of discrimination and progression in the design and specification of *Assessment Criteria* descriptors for the given programme[407]. The results are reported in Chapter 6.

The Longitudinal Dimension of Data-Collection and Processing

A third strategy of research method replicated the entire process of quantitative data-collection in cycles, as in *Action Research.* In assessing *Written Production*, cycles were completed for the May 2000, 2001, 2002 and 2003 sessions. In moderating *Internal Assessments*, cycles were completed for May 2001, 2002 and 2003. Analysis was completed by **IBO** deadlines, respecting regular criteria and procedure. However, contemporaneous experimental triangulation of the same material was only completed in moderating *Internal Assessments*, as recorded[408]. The procedural variance between oral and written assessments is indicated in Chapter 6, with discussion of likely effects in reporting research evidence.

CHAPTER SIX

EVIDENCE

Preface

For addressing research aims identified in Chapter 1, as well as key and subsidiary questions derived from these aims[409], evidence was collected and analysed as follows.

IBO Publications and Documentation for Internal Use in Formal Assessments

IBCA archives were searched for sources of evidence to supplement publications, described and discussed in Chapter 2, in partial response to all the research questions. Relevant documentation investigated includes the following[410]:

- *General Instructions* (May/November 2001) for examination production at all levels, for all languages;
- *Paper Specific Instructions: Language B, Higher and Standard Levels* (May/November 2001);
- *Checklists* for reporting and evaluating conformity to criteria in examination production;
- *Standardiser's Guidelines: Language B* (2001), (in newly-produced, draft form);

- Sampled moderation statistics for *French B* (May 2000, 2001 and 2002);
- **IBCA** correspondence between examination-designers, standardisers and production personnel, with examination drafts (*French B,* May 2001).

The collected documentation was analysed as described in Chapter 5. Relevant facts are reported, with the key aim of improving understanding of validity in standardised task-designs that promote authentic expression, albeit constrained by inflexible rubrics[411]. However, as particular interpretations arise from description and analysis, certain points will be summarised for discussion.

General Instructions for designers state that examination-design must express the philosophy, aims and assessable objectives of **IBO** programmes. Through continuous scrutiny, designers should ensure that examinations "adhere [.....] to the criteria laid down in the relevant published guide", assuming an overall consistency. *Standardisers* and *Subject Area Managers* check conformity to *Guides* and *Assessment Criteria*, commenting on proposals where necessary[412]. As referent documents, they define programmes and implementation procedures[413].

Task-Design and the Editing of Authentic Texts as Resources

As posed in the key questions for research and in a partial attempt to understand authenticity as theory in task-design, administrative and assessment practice, editing policy for 'adapting' authentically-sourced, examination reading materials[414] was scrutinised: adaptations could potentially distort written responses to *Paper 2* tasks that refer to a *Paper 1* text. In documenting examination preparations for May 2001, comments by designers reveal that 'simplifying' realia, rendering texts culturally and linguistically 'appropriate' at pre-defined levels (*Higher* or *Standard*), may influence expression through artificially directing reader attention. Although *Text-Handling* examination designs have not been extensively analysed, the effects of linking *Papers 1* and *2* requires inspection in such cases. Linkage constrains both *Written Production* task-designs and the range of 'appropriate' response, positioning organisation and examinee, creating ambiguity in understanding authenticity as theory in practice, and raising questions of validity in task authenticity, particularly as *User Authenticity* and *Authenticity of Context.*

Further practical considerations cover effects on authentic expression of potential, yet exceptional instances of examination 'malpractice', as defined by the IBO[415].

Design guidelines included in *Paper Specific Instructions* and *Checklists* for reporting and evaluating conformity to criteria, inform analysis and discussion of authenticity as theory in practice. They illuminate investigation of 'positioning', as defined by the subsidiary research questions. The concept is fundamental to language philosophy, delimiting domains for 'second', or 'foreign' languages at differing levels, and thereby 'positioning' both institution and examinees[416]. *Language B* is typified in organisational understanding and intent, thus:

- "the [.....] course is designed for students who have studied the language for between two and five years prior to the beginning of their IB course"[417];

- "the same level of sophistication cannot be expected from *Standard Level* candidates as can be expected from *Higher Level* ones"[418];
- "the format of the examination papers is the same for both *HL* and *SL* [.....]. However, the choice of questions should reflect the difference of expectations between the levels"[419].

In *Written Production,* designers should ensure that:

- "the link[ing of one *Paper 2* task] with [stimulus material in] *Paper 1* should only be tenuous in order not to disturb or frustrate candidates"[420];
- "in order not to limit a candidate's choice of written tasks to only one or two, [there are] a wide variety of questions [.....], accessible to candidates from differing backgrounds"[421];
- "the paper [.....] include[s] a variety of different tasks"[422];
- dialogues and conversations, which "can turn into artificial activities", be "avoided"[423];
- "all tasks [.....] provide a context, the type of text which is expected [as response], the audience [and] some indication of the type of register (even though it may be implicit)"[424];
- "the type of task [be] suitable to the topic and the level"[425];
- "the assessment criteria [.....] can be applied to the candidates' work. All questions should enable examiners to use all three criteria in their entirety (*Message/Task, Presentations, Language*)"[426];

- tasks [be] sensitive to the international context of **IBO** programmes and examinations: they should avoid causing "offence" in "social and political contexts which have different religious and moral beliefs and social conventions"[427];
- "questions cover a *range* of interests" and "avoid gender bias"[428];
- "questions are relevant and interesting to a 17 -18 year old student"[429];
- "Literary questions [.....] be worded in such a way that any text studied could be used to illustrate the answer. However, questions which are so general that they could be easily rehearsed beforehand must be avoided"[430].

The associated *Checklist* adds requirements that each task:

- "has been narrowed down"[431];
- "[be] meaningful"[432];
- "can be completed in 1hour, 30 minutes"[433].

For *Written Production*, a theory of authenticity is evident. The documentation promotes the avoidance of 'artificiality', the provision of variety for contrasted, contextualised task-choices, the stimulation of 'interest' and assumedly, thereby the facilitation of 'authentic' response. Regardless of personal situation, candidates may partially control presentations of 'self'. By respecting the socio-cultural, religious and political *milieux* of examinees, task-designs should allow originality and individual perspective in language-production, integrating potential readerships within the communicative processes linking 'self' with 'other'. *Creator Authenticity* and occasions for *'Personal Authentication'* are favoured in this way. Authenticity becomes "meaningful[ness]" in stimulations and assessments of contextualised, linguistic representations interlinking task and response in shared constructions of meaning. Requirements for "narrowed down" and "meaningful[ly]" situated tasks also delimit bounds for assessable production. **IBO** programme-designs specify performance ranges, not only through categorisations by level, but also through definitions of time-allocations, minimum word-length prescriptions, specific production rubrics and discrete assessment criteria.

Nonetheless, the task-setting criteria for *Written Production* are ambiguous. Instructions permit resolutions of 'inappropriacy', or 'inconsistency' through recourse to professional judgement. Although consensus-building amongst designer-groups has been neither documented as process, nor fully researched, varied patterns of thinking about authenticity emerge, meriting further consideration.

Limited by the number of focussed tasks set through designer prerogative, full autonomy of response is reduced in potential range[434]. Scope for authentic expression is partially constrained, despite the mitigation of choice provided for examinees, affording degrees of independence through its exercising[435]. From designers' perspectives, task-designs that are both 'meaningful' and 'interesting' to generations of typical candidates may further constrain performance. Individual designers' non-negotiated fields of knowledge structure understandings and interpretations, with anomalous cases of candidate response suggesting communication difficulties and requiring detailed research[436]. The balance of Van Lier's triad to which authenticity is key[437], may be adversely affected.

In contradistinction, the documentation records that design imbalances may be 'compensated' at *Standardisers'* or the *Subject Area Manager's* recommendation, with tasks remodelled accordingly.

Overall, **IBO** design instructions reveal a coherent organisational conceptualisation of authenticity, though in certain instances this is implicit as theory in practice. Assessment-tasks must not only fulfil necessary functions within constrained, examination contexts (albeit selected, defined and regulated by the **IBO**), but also stimulate language-productions as authentic responses with 'wider' concerns for self-expression. Purely psychometric or abstractly linguistic criteria are subordinate for assessment.

However, requirements defined in internal *Guidelines* do not entirely replicate those of the *Subject Guides*[438]. Rather, **IBO** instructions to *External Advisors* and *Standardisers* elaborate organisational understandings within a broader framework than published in the *Guide to Language B* for example. Through comparing these sources, internal documents emerge as more refined, precise in terminology and detailed in contextualisations[439].

For preparing examination papers, *Guidelines* and *Checklists* for *Standardisers* include additional, significant criteria[440]. Ambiguities in *Paper Specific Instructions* for *External Advisors* and examination-designers are thus partly resolved. For all examinations, a *Standardiser* must ensure:

- "conform[ity] to the same rules and regulations"[441];
- "a comparable level of challenge to candidates, irrespective of which *Language B* they study"[442];

- conformity with *Paper Specific Instructions* under which examination-designers' duties are executed, with the distinction that *Higher Level* papers "should demand a higher level of linguistic ability and sophistication" [than *Standard Level* papers][443].

In focusing *Standardisers'* attention and defining scope for recommendations, **IBO** *Instructions* and *Guidelines* propose reductions of inconsistency through 'narrowing down' and ensuring 'meaningfulness'. Hence, by triangulating the varied perspectives (of examination-designer, *Standardiser* and *Subject Area Manager*), effective balances in promoting responsive awareness, autonomy and authenticity may be maintained, and Van Lier's wholly interdependent, triadic conceptualisation of authentic communication, respected. All understandings should concur, thus obviating the need for further editing of texts or tasks.

Nonetheless, newly-drafted **IBO** *Standardiser's Guidelines* make it explicit that *Higher Level* task-responses "demand a higher level of ***linguistic ability***[444] and sophistication [than *Standard Level*]"[445]. Distinctions in "ability" are left undefined. It is unclear whether linguistic *knowledge*, *skill,* (understood as Hymesian 'competence', or appropriateness in content and in manipulating communicative structure)[446] or any combination of these, are inferrable. Comparison of *Criteria* for *Paper 2* at *Higher* and *Standard Levels* suggest such a case.

The more detailed listings of standardisation requirements by level[447] make task, rather than language-based differentiation, the predominant design-criterion, though distinctions are imprecise. *Inter alia,* examination tasks should be:

- "accessible even to weaker students, while allowing stronger students the opportunity to excel;
- appropriate to the level (*HL* questions should overall be more challenging than *SL* ones)"[448].

Discrete language levels, categorised by programme, should be further scrutinised for issues of authenticity, given ambiguities in differentiation between *Higher* and *Standard Levels*. However, the relevance is minor in this research, since only productions assessed and evaluated at *Standard Level* have been selected, described and manipulated as evidence of performance.

Documentation on the design, standardisation and production of the May 2001 *French Language B* examination[449], with related semi-structured interview of the **IBCA** *Director of Assessment* and an *Examination Papers Officer*, reveal that[450]:

- for English and French, examination-designers, or *External Advisors* on examination design, and *Standardisers* are always native-speakers;
- given 'correct' applications of design, standardisation and production criteria, longitudinal standardisation of similar examinations for validation over time is not deemed significantly meaningful;
- the entire process for producing this examination had taken fifteen months, with drafts of *Paper 2* amended twice by two **IBCA** officers, prior to proof-reading for publication;
- amendments and revisions focus on questions of grammar, vocabulary, cultural contextualisation, the 'realism' or perceived 'artificiality' of tasks proposed, and the ensuring of 'appropriate'

differentiation between demands and expectations at *Higher* and *Standard Levels*.

Whilst unresearched in themselves, designs that are securely-referenced to native-speakers' cultural and linguistic understandings, and triangulated for consensual validity, satisfy requirements for authentic contextualisation. To reduce ambiguous interpretation however, the criteria need investigation in application, with specific performances exemplifying discerned effects. Each editing category of amendment and revision may then be discretely scrutinised.

Documentation for six *Written Production* tasks to be published in May 2001 provides examples. Grammatical concerns explain editorial changes for reducing gender specificities, elegantly addressing candidates and ensuring conformity with **IBCA** requirements[451]. Vocabulary items, presumed accessible at *Standard* rather than *Higher Level*, are isolated as hypothetically difficult or confusingly ambiguous, as illustrated by the task proposed for *Question 6*.

Here, authentic advertising copy was reproduced, relating offers of *voluntary,* holiday employment ("*bénévole*" in the original French), and inviting appropriate reply. At an *External Advisor's* suggestion, the adjective "*bénévole*" was deleted from the text. For candidates and reader-assessors alike, the term could inappropriately focus, or distract attention. Anticipating requirements for specific offers of voluntary work, rather than recounting relevant employment experiences could predominate in responses, distorting the intentions and linguistic demands of the task. It could thus prove 'atypical', or 'unrealistic' in arousing interest amongst examinees, or in addressing likely experience. Hence, a broader, less constraining proposal was preferred and the task-design rephrased to omit reference to *voluntary* work. Specific authenticity was reduced, inasmuch as best productions would convince through addressing the prime intentions of original material provided for responses, requiring some discussion of voluntarism in its own right[452].

The example questions 'appropriacy' in cultural contextualisations, and authenticity as 'realism', or perceived 'artificiality' in representations proposed as tasks. For May 2001, the discussion of various design-alternatives in drafting *Question 1* provides further illustration.

This task required personal diary entries as responses reflecting on prospects of leaving a family home to live elsewhere[453]. The initial proposal created a precise scenario of leaving "parents" for "university attendance", requiring recontextualisation since its specificities might not relate to the likely experience of all examinees. Ambiguity was introduced to prevent exclusion and ensure full maintenance of choice of response, through facilitating freedom of interpretation. Hence, the term "parents" was replaced by "family", and the situation of "going to university" was deleted, without substitution[454].

Here for adolescents, editing to *External Advisors'* recommendation improves relevance and potential task-appeal at appropriate levels, relating these to likely experience in international, school and family life, and thereby addressing appropriate domains of linguistic competence.

Similar concerns typify the design, standardisation and production of *Higher Level* examinations. Archive material for the same session illustrates problems, with similar highlighting. 'Appropriately' differentiated expectations and demands of "linguistic ability" or "sophistication" at *Higher* and *Standard Levels* are at issue in such findings, although the phenomenon has not been explicitly addressed. It merits deeper investigation in future research.

The *Internal Assessment* Component

Recapitulating earlier description[455], this component comprises the assessment, moderation and evaluation of interactive performances in listening and speaking. It integrates aural and oral skills through continual, productive interchange, eschewing discrete testing at fixed times. Given necessary teacher involvement in sampling for moderation, assessment is administered by schools during the final year of a course. Hence teachers as facilitators and interlocutors, intimately share and partially shape both communicative language production and its evaluation. Analysis of the evidence collected should therefore enhance understanding of organisational conceptualisations of authenticity, addressing a key research question. It should also inform subsidiary questions through examining the specification of rubrics that 'position' the **IBO**, its *Assessors* and *Moderators,* and the candidates whose performances are analysed[456].

In this context, the programme *Guide*[457] clarifies administrative rubrics for teachers and *Internal Assessors*. These are useful as further evidence illustrating organisational conceptualisations of authenticity, outlining implementation procedures as theory in practice, and shaping performance, its recording, assessment and moderation. Constraints for 'successful' productions, with "oral work" understood as simultaneously combining oral performance with aural competence, are summarised by statements that assessments should:

- take place continuously, globally and in 'balanced' fashion, throughout the final year of the programme:
    - integrating listening with speaking;
    - permitting 'relevant' activities that "cannot be externally evaluated"[458];
    - confining 'oral work' to range of activities, as defined[459];
- under normal circumstances, be evaluated by the candidate's teacher for **IBO** moderation, according to **IBO** criteria[460];
- be recorded in writing by teacher-assessors;
- be tape-recorded in the case of individual orals between candidates and teacher-assessors, with samples sent to **IBCA** for moderation and evaluation. These must comprise orally-based, presentations of candidate-chosen, programme topics, followed by relevant discussion of six to seven minutes' duration, and concluding with more general, unprepared, yet personalised conversation of three to four minutes' duration, the whole being on average, ten to twelve minutes long[461].

Suggestions of what is appropriate stress the following:

- "All assessed activities should be related to at least one of the three course themes"[462];
- An activity should comprise a presentation based on at least one item from a candidate's curriculum dossier: for example, "pieces of writing produced by the candidate, printed texts, articles or pictures [.....], or a mixture of all these", though not exclusively so. Illustrative material and notes may be used as reference, but verbatim reading aloud is not permitted[463];
- Less-prepared, personally-based, 'spontaneous' conversation could include "the candidate's own interests (for example, books

or films the candidate has read/seen); issues affecting young people in general (for example, personal relations, education, employment); social issues (for example, crime, drugs, health); or world problems (for example, war, energy, terrorism, current affairs) [.....] Especially at *Higher Level*, conversation should go beyond the daily routine or future plans of the candidate and test his/her ability to defend opinions and counter those of another person"[464];

- The compulsory completion of at least one individual oral (comprising candidate and teacher as interlocutor), and one group, or paired oral (comprising teacher and/or at least one other student as interlocutor). Group orals comprising at least four candidates are preferred, where feasible[465];
- "At least one of the assessed activities should be based on a listening stimulus [.....] in such a way as to make it possible to apply the internal assessment oral descriptors. For example, candidates may be asked to watch a film in the target language and then discuss their impressions"[466];
- "At *Higher Level*, at least one of the assessed oral activities must be based on literature [.....] Suitable activities [.....] might include: oral commentary on an extract from a work studied as a part of the programme; discussion on a particular aspect of a writer's work; a presentation of a comparison of two passages/two characters/two works; a role play dialogue between two characters from different works/from the same work, discussing their contrasting motivations, explaining their behaviour, etc.; a role play interview of an author by one of his/her characters; a role play interview of a character from a work of fiction interviewed by a candidate either as him/herself or in another role (such as a psychiatrist or a social worker)". It is

emphasised that "this list is neither exhaustive nor compulsory. Some examples may be inappropriate to some *Languages B*"[467];

- The above requirements may be combined in a single activity[468].

Notions of authenticity underlie such rubrics. Indeed, the *Internal Assessment* design appears predicated on conceptualisations presented and discussed earlier. The selection, specification and creation of situated *content* must stimulate interchange, facilitating interdependency and integration in listening and speaking, thus providing evidence of authentic communication. Secondly, the choice, comprehension, completion, assessment and evaluation of contextualised *tasks* through which content is formalised, must be grounded, both linguistically and socio-culturally, in 'meaningfully' interactive language use.

The requirement that candidates present topics *of their own choice*, reflecting the *interests of 'self'*, personally selected from broadly-generalised, **IBO**-defined fields, and relating to individual, candidate-based 'research', recorded in curriculum dossiers of materials studied in the course of the programme, satisfies criteria for authenticity. They are those defined under categorisations of C*reator Authenticity as 'self'*, *Finder* and *User Authenticity*, as presented in Chapter 5[469].

Authentic performance should also be facilitated in distinctive productions by the requirement that presentations and discussions integrate listening with speaking in a 'balanced' way[470], and that this balance include the spontaneous development of linguistically-embedded, socio-cultural and communicative relationships between candidates and teacher-assessor-interlocutors. Here, Van Lier's categories of *Creator Authenticity as 'other', User Authenticity, Authenticity of Context,* of *Purpose,* and of *Interaction* are significant.

Productions according to rubric may simultaneously indicate evidence for the more psychologically holistic categories of *Existential, Intrinsic* and *Autotelic Authenticity*, in their triadic relationship with candidate awareness and autonomy, as presented in Chapters 3 and 5.

Criterion Descriptors for *Internal Assessment*

Authenticity serves as a working concept for designing and implementing **IBO** *Group 2 Languages* programmes. Equally, it is a referent for assessing the processes and outcomes of resultant language-production[471]. Under the relevant criteria, highest-scoring oral performances are therefore judged:

- "interesting; comprehensible; clear; coherent; relevant to the topic chosen; convincing and in part original", in *Message* and as *Task* response*;*
- "lively; actively participatory in discussion; fluent; sensitive and nuanced", as *Interaction;*
- "fluent; varied and largely correct, with expressive intonation and pronunciation that facilitates communication", in *Language.*

In contrast, the lowest scoring, and hence least authentically-viable performances are judged:

- "extremely superficial or incomprehensible; repetitive and/or irrelevant", in *Task* and *Message;*

- "lacking in coherence; reluctant; in need of prompting; limited in comprehension; inconsequential", in *Interaction;*
- "generally incomprehensible; limited in range; inexact, grammatically incorrect; with intonation and pronunciation that impede communication", in *Language.*

Understandings, choices of oral topic, responses to tasks in presentation and discussion, teacher-assessor interventions, **IBO** criterion-descriptors and assessment procedures as applied by *Assessors* and *Moderators*, all come into play, globally positioning candidates in a constrained, yet 'appropriate' way. Summarised in the previous section, practical administrative considerations also form rationales for additional constraint[472].

Preliminary analysis of *Internal Assessment* data for better understanding 'problematic' responses to conceptual inconsistency within the programme[473] shows that under the **IBO** criteria, certain productions inferentially render authentic expression difficult. For further discussion, the programme constraints require detailed scrutiny. Practical limitations in conceptualising authenticity as a design-aid for specifying high-validity assessment regimes can be easily detected. Indeed, certain anomalous cases, illustrating assumedly 'inauthentic' task-response and intervention by 'others', affecting performance, are described and analysed in subsequent sections of this chapter. However, inasmuch as **IBO** programmes 'impose' non-negotiable restrictions in style and range of response, hence constraining fully authentic performance (at least theoretically so), effects in most cases are neither evident nor easy to measure.

Paper 2: Written Production

Internal Assessment has been contrasted with the assessment of *Written Production* for identifying theory and practice, integrating the *Diploma Programme* as pedagogy and learning, with the assessment and evaluation of its products: a key aim of the research.

To recapitulate, the writing component comprises production, assessment, moderation and evaluation of interactively-produced language[474], though with less interaction than for *Internal Assessment*, candidate competence being tested in individual examination[475]. It unifies the partially-interdependent skills of reading and writing. A short reading component comprises a task-specification, stimulating written response (though some examples of task-requirements, such as dramatisations, dialogues, speeches and so forth, suggest simulation of listening and speaking[476]). Accurate comprehension is therefore a pre-requisite for appropriate response.

For *Paper 2*, the linguistic perspectives and contributions of others merely initiate interchanges between reader and writer. Candidates control performance, albeit under set rubrics and task-requirements. Assessment and evaluation are completed independently, fundamentally differentiating the processes from those for *Internal Assessment.*

IBO statements on the *Nature of the Subject* specify linguistic interaction as "communicative"[477], with assessment criteria focussing "principally on interaction between speakers and writers of the target language"[478]. This significant aim promotes productive and situated language use, within contexts defined in published *Objectives* as "social", "academic" and "cultural"[479]. Recommended pedagogy in preparing students for *written* performance, is the study of "a wide range of oral and written texts of different styles and registers", with recourse to "authentic materials [.....] wherever possible"[480].

For enhancing motivation and commitment, teachers should encourage student participation in the selection of topics and texts[481]. In terms of pedagogy and learning, key conditions for authentic communication are facilitated by group negotiations of choice. The design of *Paper 2* provides opportunities for candidate-choice, though limiting the selection to one task from six. Issues of standardisation and equity in determining task-equivalence are raised, albeit without restricting choice as an authentic operation in itself. Tasks are neither weighted, nor differentially-assessed to account for relative ease or difficulty[482].

The programme aims define appropriate contexts for task-based, authentic expression, promoting in reiteration:

- accurate and effective communication with others through target language use in speech and writing;
- transactionally and socially-contextualised communication;
- learning effectively applying to employment or leisure-time activity, and allowing the pursuit of study interests;
- learning that integrates language with "insights into the culture of the countries where the language is spoken";
- individual motivation through opportunity for "enjoyment, creativity and intellectual stimulation"[483].

In situated, oral task-design, the specification, selection and creation of *content* should stimulate and facilitate continuous authentic expression.

Choice, comprehension and the completion of tasks for processing this content must also be grounded in communication that is to some degree linguistically and socio-culturally appropriate, hence meeting interlocutor expectations[484].

However, *Paper 2* is clearly distinguished from *Internal Assessment* in requiring responses that respect well-defined discourse genres, according to task-choice. Performance is precisely situated by prescription. Candidates may not present topics *of their own choice*, to reflect individual 'self'-expression in their own fashion, but must respond to the expectations of 'other', through producing authentic texts.

Whilst the design for *Paper 2* partially satisfies criteria for C*reator Authenticity as 'Other'*, *Finder* and *User Authenticity*[485], performance constraints may hinder displays of competence in providing evidence for awareness and autonomy in comprehension. Possibilities for negotiation are limited and the production of fully authentic responses thereby inevitably restricted[486].

Additional requirements that written, task-based productions situate chosen responses within appropriate cultural contexts in a 'convincing' way, demonstrating awareness of the likely expectations of specified readerships[487], do not impair the potential for authentic performance. Here, *Creator Authenticity as 'Other', User Authenticity, Authenticity of Context',* of *Purpose,* and of *Interaction* are relevant. Assessors should recognise the readership addressed, and consistently judge from an appropriate perspective. In contradistinction to real-time evolution in *Internal Assessment* performances, *Authenticity of Interaction* is likely however, to be displayed merely through initiating written responses, through preliminary reader involvement with a chosen task.

Nonetheless, certain compositional genres (such as theatre, dialogue and formal speeches) do indeed allow demonstration of authentic, linguistic 'interactivity', albeit internally within text-productions and constrained as inflexible through 'fixing' in writing.

As overall with *Internal Assessment,* the design for *Written Production* may reflectively supply evidence for the more psychological categorisations of *Existential, Intrinsic* and *Autotelic Authenticity.*

Criterion Descriptors for *Paper 2: Written Production*

Given dissimilar 'expectations' between *Standard* and *Higher Level* examinations and assessment[488], implicit in *General Instructions*, *Paper Specific Instructions*, *Checklists* for *External Advisors* and task-designers, and more explicit in draft *Guidelines* for standardisers, **IBO** use of little-defined, yet key concepts of differentiated "linguistic ability" and "sophistication" are better understood from close analysis of assessment-criterion descriptors for *Written Production.* This satisfies research aims of illuminating authenticity as an all-embracing concept guiding programme design and implementation[489].

Terminology is more precisely apprehended through horizontal comparison of *Level* descriptors, across different, yet like-scored criterion-categories, and vertical comparison of descriptors, only differing in value within a single level and criterion. Situated meanings and specific usages better characterise "linguistic ability" and "sophistication", refining conceptualisations of authenticity as **IBO** theory in practice.

In addition, both criteria and processes for transforming a range of qualitatively-categorised assessments into individually-aggregated, quantitative scores that are finally transformed into overall grade evaluations[490], are more clearly delineated and opened for critical analysis and appreciation.

The content of general descriptors may be tabulated, with tables devoted to each assessment criterion. This facilitates comparison across point-value categories and across levels, as subsequently shown[491].

Distinctions in interpreting quality *gradations* are highlighted in blue, whereas those for qualitative *categorisations* are highlighted in red. Descriptions in blue are compared in intensity, whereas those in red are compared by conceptualisation, with bold-type indicating distinctions requiring little further analysis, and standard-type indicating ambiguities more freely open to subjective interpretation by individual assessors.

Shown overleaf, minimal production at either *Standard* or *Higher Level*, is undifferentiated and evaluated at zero. However, identical performances are unlikely to have been elicited[492].

Furthermore, the evident repetition of similar descriptors across levels implies the achievement of *Level* differentiation solely through task-specification.

Table 6.1

Criterion A Descriptors: ***Task and Message***

Point Value[493]	***Standard Level***	***Higher Level***
Zero	No given descriptor applies	
1/2	Task completion **generally inadequate**. Message frequently **incomprehensible**.	Task completion **on the bare limits of adequacy**. Message frequently **unclear**
3/4	Task completion on the bare limits of adequacy. Message occasionally incomprehensible.	Task completed only at a superficial level. (The candidate never goes beyond the obvious in the terms of the task). Message sometimes unclear.
5/6	Task appropriately (or adequately) completed. Message **generally comprehensible**.	Task appropriately (or adequately) completed. Message **comprehensible**.
7/8	Task **generally completed well**. Message comprehensible.	Task **completed well**. Message comprehensible **and interestingly presented**.
9/10	Task completed **well**. Message comprehensible and interestingly presented.	Task completed **very well**. Message **attractively**, interestingly and clearly presented.

At *Standard Level* for *Task*, point-values per discrete category rise accordingly:

> "[inadequate] > generally inadequate > on the bare limits of adequacy > appropriately (adequately) completed > generally well completed > well completed".

At *Higher Level*, the comparable progression is:

"[inadequate] > on the bare limits of adequacy > superficially completed
> appropriately completed > well completed >
very well completed"

As colour-indexing illustrates, there is clear progression across *Standard Level* descriptors, within a single, over-arching conceptualisation of 'adequacy', or 'appropriacy', whereas for *Higher Level*, progression from "the bare limits of adequacy", (or in official English versions: "barely adequately carried out"), to being "superficially completed" (or "never go[ing] beyond the obvious") is interpretatively more ambiguous[494]. Crucially, the distinction blurs boundaries between uncontentiously 'inadequate' performance valued at 1/2 points, and performance close to the significant 4/5 score boundary, valued at 3/4 points[495].

At *Standard Level* for *Message*, point-values per discrete category rise similarly, as follows:

"[incomprehensible] > frequently incomprehensible >
occasionally incomprehensible > generally comprehensible >
comprehensible > comprehensible and interesting"

In comparison, those for *Higher Level* progress thus:

"[unclear] > frequently unclear > sometimes unclear >
comprehensible > comprehensible and interesting >
attractive, interesting and clear"

Once more, there is ambiguity in the progression from an implied ‘unclear’, evaluated at zero, to “frequently unclear”, or from “sometimes unclear” to “comprehensible”, with detailed descriptions affording little further clarification. Differentiation appears based in frequency of occurrence of evidence, with the notion undefined and left to assessor or moderator judgement. However, “sometimes unclear” also means that “the message or arguments are *barely convincing*”, or are “*on the limits of being convincing*”[496], whereas “comprehensible” refers to message or sets of arguments that are “partially convincing”. It is noteworthy that an additional criterion is described in detail, distinguishing categories valued at 3/4 points and 5/6 points, respectively. Under the former, candidates are described as “making little attempt to respond to the expectations of readers”, whilst under the latter, they should provide evidence of “making a clear attempt to respond [thus]”[497]. However, point-value boundaries remain blurred by criterion-descriptions requiring readers and assessors partly to infer authorial intention from the textual evidence presented.

Ambiguous differential intensities and inconsistent criterion-descriptions confuse the demarcation of assessment categories by six discrete, equally-weighted, point-value groupings in all *Group 2 Languages* designs. Categorising descriptors by conceptualisation and intensity, and eliminating to the maximum degree possible, scope for either ambiguity or inferential interpretation of candidates’ states of mind[498], require neither discrete groupings by six, nor balance in point-value weighting within each category. Hence **IBO** policy for symmetrically-common design, with assessors measuring each criterion by selecting one of six descriptions, equalised across all three criteria, may derive more from administrative needs for quantitative evaluation than from any requirement for qualitative, and thus more authentic assessment.

Indeed, the *a priori* significance of 'symmetry' in quantitative designs, numerically valuing qualitative descriptions without reference to interactivity in linguistic relations between 'self' and 'other', is irrelevant to measuring quality in communication. Issues of validity and reliability therefore inevitably arise.

A similar exercise (with similar privileging of original French-version sources, cross-referenced to official, English-language versions) is revealing, when applied to the *Criteria* for *Presentation* and *Language*, tabulated as shown overleaf

Table 6.2

Criterion B Descriptors: ***Presentation***

Point Value[499]	***Standard Level***	***Higher Level***
Zero	No given descriptor applies	
1/2	Presentation **poor and unclear**. **No apparent (or no attempt at)** structure.	Presentation **poor and unclear**. **No apparent (or obvious)** structure.
3/4	Presentation occasionally clear. **No clearly apparent (or no real attempt at)** structure.	Presentation barely effective. **Little apparent** structuring
5/6	Presentation generally clear, yet with faults (or occasional lapses). Attempts at structuring have been made.	Presentation reasonably effective. Structuring reasonably effective.
7/8	Presentation clear. Good attempts at structuring have been made.	Presentation effective. Clearly apparent structuring.
9/10	Presentation **effective**. Structure clear.	Presentation **inventive (or imaginative) and effective**. Well balanced structuring.

Table 6.3:

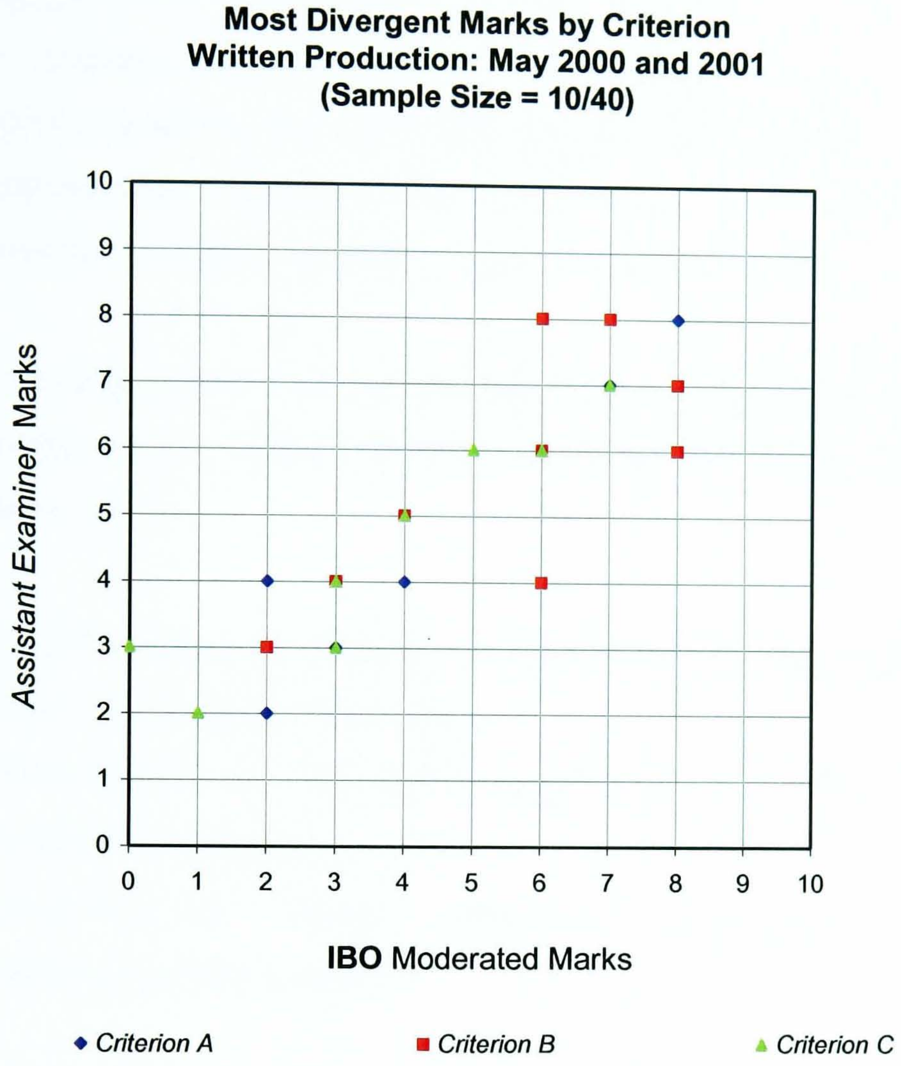
Criterion C Descriptors: *Language*[500]

Point Value[501]	***Standard Level*** **(from French descriptors)**	***Standard Level*** **(English descriptors)**	***Higher Level*** **(from French descriptors)**	***Higher Level*** **(English descriptors)**
Zero	No given descriptor applies			
1 or 2	**Overall** use of language **incomprehensible**.	Language **on the whole not comprehensible.**	Language **comprehensible**, but **clumsy and inappropriate overall**. Style **awkward.**	Language on the whole **laboured, inaccurate and lacking in fluency**
3 or 4	Language not always **comprehensible.** There is an **awkwardness** in style.	Language **not always comprehensible** and **lacks fluency**.	Language limited **overall**. There is **sometimes** an **awkwardness** in style.	Unambitious language **on the whole**, with **some lapses** In **fluency**.
5 or 6	Language **comprehensible** overall.	Language **on the whole comprehensible.**	Overall language use and style competent and **fluent**.	Language mostly **fluent**.
7 or 8	Language use generally competent and **fluent**.	Language mostly **fluent.**	Language use competent and **fluent**.	Language **fluent**.
9 or 10	Language use creates an impression of competence, fluency and authenticity, thereby diminishing the seriousness of any errors that arise.	Language **fluent**. Natural ring reduces the impact of any minor mistakes.	Language use competent and **fluent**, creating an impression of authenticity.	The language **fluent with an authentic ring**.

After analysis, certain significant criteria[502] appear categorised in a way requiring further investigation. Under the *Criteria* illustrated by these tabulations, rationales for descriptions and numerical evaluations remain implicit. Ambiguities occur in different versions, and differentiation in gradations of performance is often imprecise. For example, *Subject Reports* make no reference under *Presentation,* to legibility of candidate handwriting with minimal deletions of work. The criterion appears subsumed under *Language,* where spelling and handwriting should be assessed for legibility and the degree to which they "disturb" matching to other descriptors[503]. Under no rubric is punctuation mentioned as a discrete category and criterion for assessment, and so forth[504].

When examined in detail, the criterion-descriptors for *Language*, reveal a linear conception of value, rising from simple usage at low levels, to sophisticated or complex expression at the highest. From assessment experience, this may hinder determinations of validity and reliability, should the criteria be strictly respected. Certain responses show that candidates may 'succeed' in producing linguistic sophistication and complexity, whilst remaining incorrect in elementary usage, creating a dilemma for matching aggregated scores to the *General Criteria* at any given grade level. For example, linguistically erroneous work may to some extent communicate messages that are clear, if inadequate responses to tasks set. Across the three criteria, major score variations may be aggregated for quantitative 'success' that is difficult to justify as an overall outcome in quality. The most extreme divergences have been isolated from **IBCA** moderation samples provided by the researcher as *Assistant Examiner*, with results represented graphically, as follows.

Figure 6.1



The score-distribution illustrates significant variation in assessor verdicts. For certain cases, the aggregated totals imply significant difference in final grade-awards. (Individual examples are subsequently described in detail). The very structuring of assessment by equally-valued and weighted categorisations of *Task/Message, Presentation* and *Language* according to implicit rationales, is thus demonstrably problematic. When matching criteria to any general philosophy of language-production, and measuring authenticity in language use, as espoused by the **IBO**, it seems that the lower the quality of performance under *Criterion C: Language*, the less the possibility should exist for appropriately-structured, functional texts. The boundaries between the **IBO**'s tripartite, assessment categories are blurred by qualitative transferences from *Language* to *Task, Message* and *Presentation*, thereby creating duplications for quantitative scoring. Identical performance qualities may be 'rewarded' more than once. Language-based criterion-descriptors for "comprehensibility" and "clarity", isolated from *Criteria A* and *B* in *Tables 6.1* and *6.2*, imply as much[505].

Discrete categorisations of components of competence further illustrate the anomaly, under definitions given in the *Principal Characteristics of the Criteria,* as[506]:

- *Task/Message* or: *Overall Competence, Content, Conformity to Task, Capacity for Argument;*
- *Presentation* or: *Overall Presentation, Structure, Cohesion, Register* and *Style;*
- *Language* or: *Overall Impression, Grammatical Precision, Vocabulary,* and *Legibility;*

Each sub-category appears conceptually interlinked and all are mutually influencing. For example: "overall competence" clearly includes attainment in *Language*; "capacity for argument" implies demonstrable ability to structure work persuasively, thus relying on aspects of *Presentation*; questions of "register" and "style", evidently relate to standards of "grammatical precision" and the appropriate choice of vocabulary, assessed under *Language.*

For exploring reliability in correlating qualitative, criterion-referenced assessments with quantitative evaluations, the problems require further description, analysis and discussion, relating data to *Internal*

Assessment Moderator and *Assistant Examiner* communications as interpretations of criteria and instructions for procedure. The following section therefore considers documentary data, published for internal use by **IBO** employees in fulfilling their duties. French language versions were supplied to the researcher as *Internal Assessment Moderator* and *Assistant Examiner*[507].

The Examiners' Handbook: Examination Sessions for May and November (**IBO** 2001a, 2002d) contains the following sections, relevant in detailing **IBO** 'espoused theory' for subsequent discussion in this chapter:

- *Second Part Section A: Receipt and Marking of Examination Material;*
- *Second Part Section B: The Composition of Reports;*
- *Fourth Part: General Instructions for the Moderation of the Internal Assessment;*
- *Fourth Part: Language B Higher and Standard Levels: Instructions for the Internal Assessment;*
- *Fifth Part: Marking Examination Material;*
- *Fifth Part: Language B Higher and Standard Levels: Paper 2, Written Production.*

For *Written Production*, the **IBO**'s assessment, moderation and evaluation 'theory in practice' at *Grade Award Meetings* may partially be found in the *Report on Attendance at the Moderation Meeting for French, Language B: November 2000 examining session*, and the similar report for *German Language B* of June 2001[508]. Comparing these data, as included in the research, illuminates theoretical frameworks, both explicit and implicit, governing situated understanding.

Reiteration serves as recapitulation for further analysis and discussion of the relevant elements.

Supplementary Documentation

This concerns procedures and outcomes for assessment sessions in *French, Language B*, and serves to improve understanding of reliability in correlating qualitative, criterion-referenced, **IBO** assessments with quantitative evaluation, together with the 'positioning' of examiners in the process.

From May 2000 to May 2001, *Assistant Examiners* retained in **IBO** employment have received commented exemplars of evaluated *Written Productions*, as feedback on their assessment practice. Copies received in February 2001 and May 2002[509] as samples of the researcher's examining performance, are described and analysed in this section.

The documentation consists of:

- a covering letter from the **IBO** *Director of Assessment*, with acknowledgement of re-employment for the ensuing session in the case of the February 2001 communication, and with summary restatement of the knowledge necessary for satisfactory accomplishment of examining duties. The letter explains selection procedures for sampling scripts from named examinees, including their grade status, and requires confidentiality in use and ultimate disposal;
- three copies of **IBCA** *pro-formae,* completed by the *French B Chief Examiner*, commenting the quality of *Assistant Examiner*

Reports that summarise and evaluate overall candidate-production from three separate centres, identified solely by **IBO** code numbers[510];

- ten copies of sampled work, with five from the May 2000 session and five from 2001, including coversheets completed by the *Paper 2 Moderator.*

The circular letters state the **IBO**'s intention to facilitate "collaboration" between members of international teams of assessors, scattered across the world. Before each session, examiners will receive comment on their assessment performance, supplementary curricular information detailing the basis of examination rubric and assessment-task design, and guidance from *Chief Examiners* on applying the *Assessment Criteria*. They outline procedures for sampling *Assistant Examiner*-assessed scripts, with annotations, re-assessment and further remarks by *Chief Examiners.* Recipients are urged to review this documentation, and note particular cases where *Chief Examiner* judgements differ from that of the *Assistant* (and hence the researcher). However, it is also explained that the selections illustrate cases of widest divergence. Individual examiner's future standards may therefore need no alteration to achieve greater conformity. Unselected copies may thus be presumed to reflect greater unanimity of assessor, moderator and evaluator verdicts. In conclusion, *Assistant Examiners* are encouraged to reflect upon future practice in the light of such feedback, aiming to maintain valid, reliable and equitable assessment and evaluation practices for all examinees. Issues of confidentiality in the use and disposal of forwarded material are reiterated[511].

It should be noted that in this context, *Team Leader* assessments and evaluations are remoderated. Reliability co-efficients for adjusting irregular scores are redetermined by the **IBO**, with *Chief Examiners* replicating the entire process.

The present *Assistant Examiner's* three *Reports* for May 2000 and three further *Reports* for May 2001, relating performance for centres requesting an *Individual School Report* (or ***ISR***), were described as "excellent", "encouraging [for the recipients]", with "good advice" and being "very complete and detailed, very useful", if on occasion "too long" and in places, "repetitive", with some "slightly contradictory remarks[512].

In scoring by component-criterion, the samples record divergences, tabulated overleaf, with candidates listed randomly and anonymously.

Table 6.4

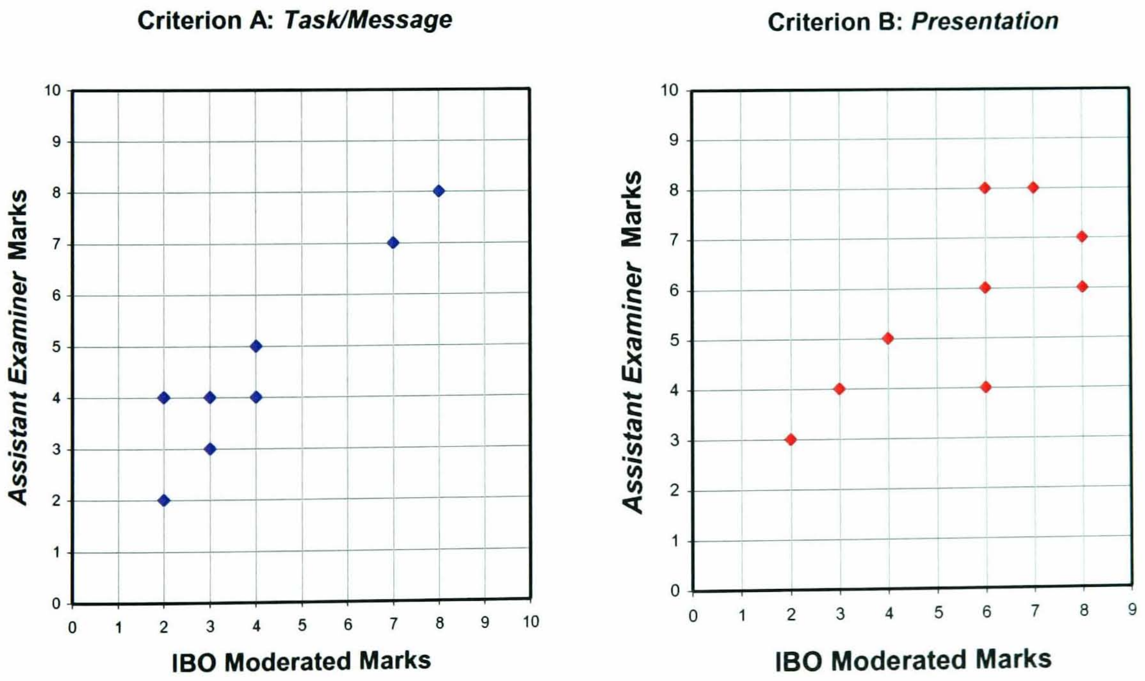
French Language B Score Divergences *Paper 2: Written Production*

Candidate	Assessment Criterion	*Assistant Examiner's* Scoring	*Moderator's* Scoring	Score Divergence	Diploma Grade Equivalence	
May 2000 Session						
1	*Task/Message*	4	3	+1	**3**	**3**
	Presentation	4	3	+1		
	Language	4	4	0		
	Total	**12**	**10**	**+2**		
2	*Task/Message*	7	7	0	**5**	**5**
	Presentation	6*	8*	-2		
	Language	6	6	0		
	Total	**19**	**21**	**-2**		
3	*Task/Message*	2	2	0	**2**	**2**
	Presentation	3*	2*	+1		
	Language	2	1	+1		
	Total	**7**	**5**	**+2**		
4	*Task/Message*	3	3	0	**3**	**3**
	Presentation	4*	6*	-2		
	Language	3	3	0		
	Total	**10**	**12**	**-2**		
5	*Task/Message*	4*	2*	+2	**4**	**3**
	Presentation	6	6	0		
	Language	5*	4*	+1		
	Total	**15****	**12****	**+3**		
May 2001 Session						
6	*Task/Message*	7	7	0	**5**	**5**
	Presentation	7	8	-1		
	Language	8	8	0		
	Total	**22**	**23**	**-1**		
7	*Task/Message*	7	7	0	**5**	**5**
	Presentation	8	7	+1		
	Language	4	4	0		
	Total	**19**	**18**	**+1**		
8	*Task/Message*	6	5	+1	**5**	**4**
	Presentation	8*	6*	+2		
	Language	4	4	0		
	Total	**18****	**15****	**+3**		
9	*Task/Message*	3*	0*	+3	**3**	**3**
	Presentation	5*	4*	+1		
	Language	5*	4*	+1		
	Total	**13**	**8**	**+5**		
10	*Task/Message*	3	3	0	**3**	**3**
	Presentation	5*	4*	+1		
	Language	4	3	+1		
	Total	**12**	**10**	**+2**		

Aggregated in *Figure 6.1*[513], this quantitative evidence has been represented by scatter graphs comparing **IBO** *Moderator* and *Assistant Examiner* marks per criterion. (The data represent examples of greatest variance from **IBCA** moderations). For each examining session, they are contrasted with similar representations of *Assistant Examiner* and *Team Leader* scorings for complete moderation samples of twenty copies, yet for which no **IBCA** report is included, (variance being less significant in the remaining cases). It is apparent that greatest variance, indicating least consensus, occurs in *Presentation* in the upper range.

Figure 6.2

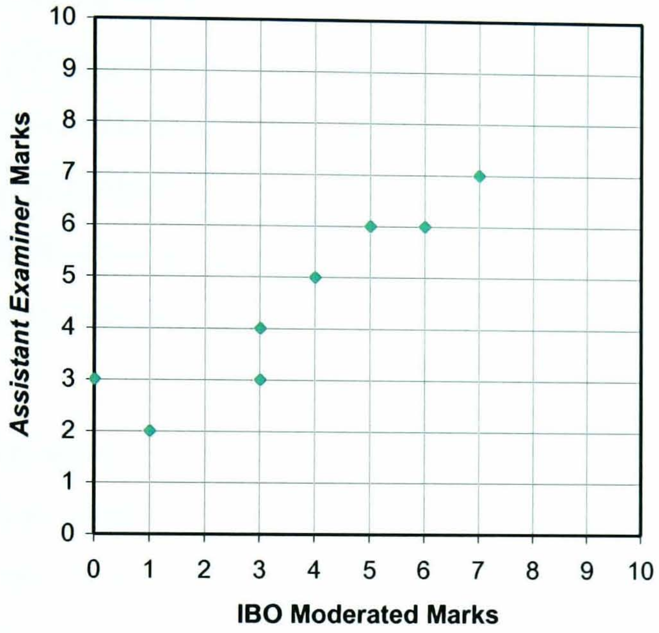
Criteria A and B: Variance between
***Assistant Examiner* and IBO *Moderator* Marks**



Least variance appears in *Language,* as shown below:

Figure 6.3

Criterion C: *Language*



This suggests that the most positivistic assessment domain (namely *Language*) allows for greatest consistency, albeit from evidence that only relates to the most divergent cases, sampled from forty written responses to twelve different tasks over two examining sessions. Assessment under *Task, Message* and *Presentation* displays greater inconsistency between assessors, albeit in a limited number of cases. For generalisation, the base of ten is small and therefore to be used with caution. Yet as Gipps claims, following Linn, Dunbar and others:

> "the evidence is that performance on performance assessment-tasks is highly task-specific; that is, performance on different tasks from the same domain, or on tasks that appear to be similar, will only be moderately related. The actual task set leads to variability in performance."[514]

In this respect, the limited finding accords with earlier research.

The same evidence is represented numerically in *Table 6.4*[515], where single asterisks indicate scores from different assessment categories, and double asterisks indicate aggregations, differing by one grade in **IBO** seven-point scale evaluations[516]. Scores within a single descriptor category are differentiated at assessor discretion, and may be manipulated for ‘compensations’. That is, ‘severity’ in one criterion, may be balanced by justifiable ‘generosity’ in another. Given this leeway, authorised by **IBO** procedure and advice to *Assistant Examiners*, such scores have not been highlighted.

From these observations and tabulations of data, two cases of variance between *Assistant Examiner* and *Moderator* scorings result in different final grades. In *Table 6.4,* these examples, (Nos. 5 and 8) are potentially significant insofar as for the former, the traditionally-accepted boundary between Grades 3 and 4 is crossed. For many ‘high stakes’ purposes this is understood as ‘fail’ or ‘pass’[517]. With the latter, the boundary between ‘average’ and good performance, respectively evaluated as Grades 4 and 5, is crossed[518].

Also noteworthy is example No. 9, awarded zero by the *Moderator* and justified as “not being in the form of a diary, as required in the task”[519]. Here, judgements could prove critical since the combination of component scores, representing the extremes of point-values for Grade 3[520], shows significant divergence, even though the verdict of the *Assistant Examiner* is generally deemed ‘reliable’.

With examples Nos. 5 and 9, significant variance relates mainly to *Task* and *Message.* For No. 8, it centres on *Presentation.* In no case is it significant in *Language.*

The greatest variation (occurring in May 2001) represents a 5-point difference between *Assistant Examiner* and *Moderator* verdicts, yet results in an invariable grade, being within the range established for Grade 3. However, if rated according to values determined under common rubrics, procedures and levels of language use for similar task completions for May 2000, (where it may be assumed that no alterations in grade-values are required, task-difficulty being equivalent), the resultant grades would have varied between 3, as determined by the *Assistant Examiner*, and 2, as awarded by the *Moderator*. If reproduced throughout the system, this outcome would be significant, not only for the subject component but also for *Diploma* awards, since gradings at 2 imperil these, requiring compensation with higher minimum grades attained elsewhere[521].

In this single, anonymous case however, it is unknown whether the total score aggregation, combining internal and examination assessments, perpetuates variance between the assessors and moderators of each component. The ultimate significance of the evaluation is therefore unknown.

Nevertheless under *Grade Award Meeting* procedure, totalised scores are directly related to overall descriptors, published in the *General Criteria for Grade Awards* and reproduced overleaf[522].

Table 6.5

General Descriptors for Final Diploma Grade Awards

Point Score		**Grade/ Summary Description**	**General Description (English version)**
May 2000	**May 2001**		
0–4	**0-4**	1: Very Poor	Incomprehensible
5–9	**5-8**	2: Poor	Makes little sense
10–13	**9-12**	3: Mediocre	Often unclear: difficult to understand. Very limited, often inaccurate vocabulary; poor grasp of grammar.
14-17	**13-17**	4: Satisfactory	Generally comprehensible: limited but fairly accurate vocabulary; frequent basic grammar mistakes.
18–22	**18-22**	5: Good	Always comprehensible but ideas are commonplace; some structure. Limited vocabulary but some idiomatic expressions; basic grammar usually correct.
23–26	**23-27**	6: Very Good	Fairly competent. Some originality. Structured clearly. Good variety of vocabulary and idiom; a variety of grammar, generally well-handled.
27–30	**28-30**	7: Excellent	Competent and generally accurate language. Original and/or convincing ideas. Clear structure with conclusions. Good range in vocabulary and idiom; flaws in expression do not obscure meaning.

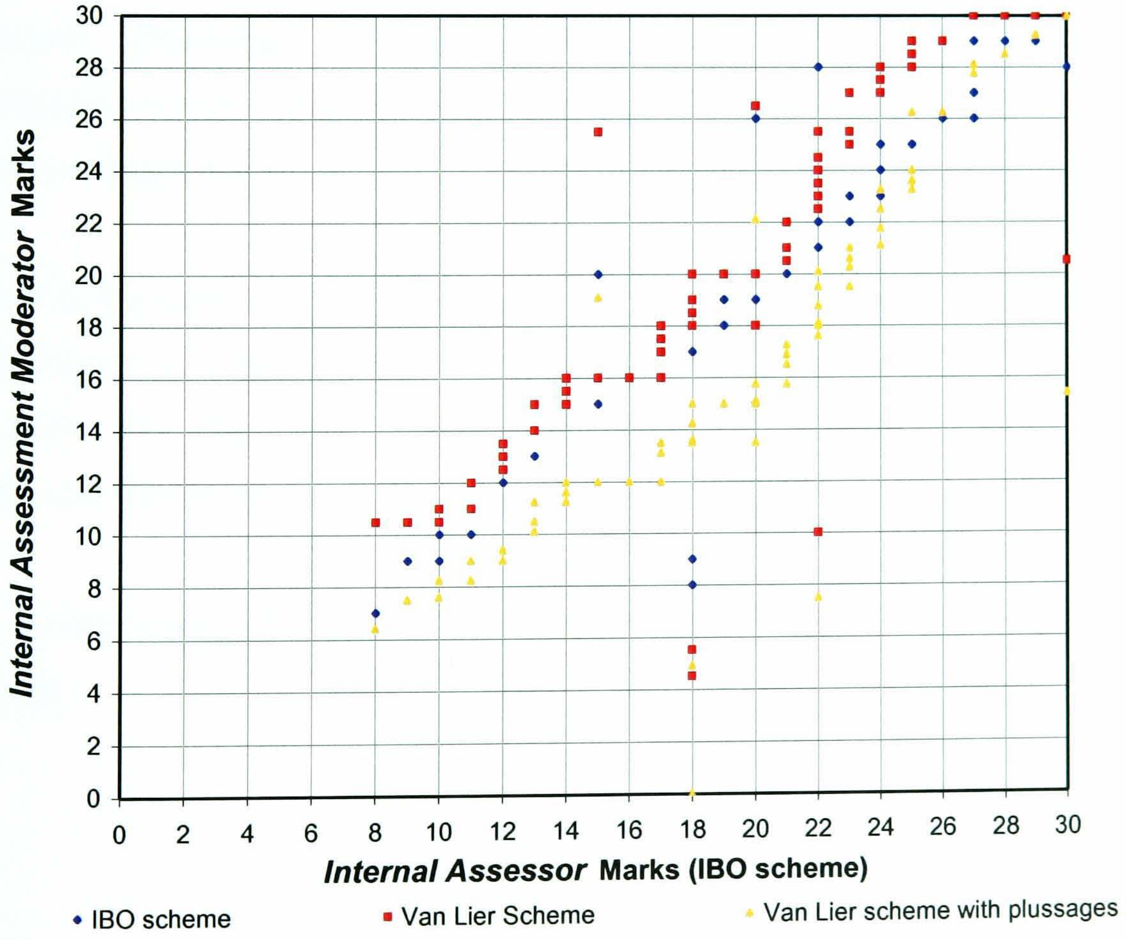
Such recategorisations, with predominantly linguistic descriptions of criterion content, appear to bias final evaluation towards purely structural, linguistic and positivistic normalisations that may be established without reference to communicative value. Concomitantly, they imply devaluations of value-weightings for *Task/Message* and *Presentation.* The categorisation of "very poor" thus relates to "incomprehensibility", "poor" to "mak[ing] little sense, and "mediocre" to being "difficult to understand". The higher categorisations of "satisfactory", "good", "very good" and "excellent" relate more to linguistic quality, with explicit, qualitative criteria for *Task/Message* and *Presentation* only apparent in productions described as "good" or better, through graduated references to "ideas [being] commonplace: some structure", "some originality: structured clearly", and "original and/or convincing ideas: clear structure with a conclusion".

In the single, significant case sampled, the *Assistant Examiner*'s verdict, quantified at 15 points, translates as "satisfactory: generally comprehensible: [with] limited but fairly accurate vocabulary; [and] frequent basic grammar mistakes". The same written production, quantified at 12 points by *Chief Examiner* and *Moderator*, translates as "mediocre" or "often unclear: difficult to understand; [with] very limited, often inaccurate vocabulary; [and] poor grasp of grammar". From tabulated comments elaborating rationales and justifying component point-awards (summarised in *Tables 6.6* and *6.7),* greatest variation appears in assessing *Task and Message.* Here, examiner concern centres on relevance to *Task,* even though the communication of message appears successful. Indeed, as recorded in analysis of ten, **IBO**-sampled scripts, agreement in judging language-quality is often close, with overall variation limited by a tendency to slight, though insignificant generosity in *Assistant Examiner* scores. From inspecting final evaluations, it may nevertheless be concluded that **IBO** grade-weightings emphasising purely linguistic criteria have no appreciable outcome.

For *Internal Assessments*, assessor divergence has been represented graphically. 100 oral presentations and interviews have been analysed for comparison of a range of *Internal Assessor* marks and those for the present researcher as *Internal Assessment Moderator*. Experimental re-assessment according to Van Lier-based criteria are also represented, as follows:

Figure 6.4

Comparison of Aggregated Marks for *Internal Assessment*: May 2001 and 2002
(Sample Size = 100)



In these cases, correlations between scores aggregating *Task/Message*, *Interaction* and *Language* seem clear and consistent. Relatively few examples are anomalous[523].

The results appear to substantiate the claim that overall outcomes per candidate are little affected, regardless of assessment scheme, be it biased towards structurally-linguistic, positivistic measurements, as with the **IBO** *Language* criteria, or towards assessing authentic communication, as under experimental triangulation with Van Lier's criteria.

Examiner and *Moderator* comment, required for justifying awarded scores, has been tabulated for *Written Production*. This facilitates comparison, candidate identification being cross-referenced to *Table 6.4*. With scorings, areas shaded lightest grey highlight greatest *Assistant Examiner* and *Moderator* variance; mid-grey indicates permissible minor variance within a single descriptor category, under assessor 'impression'; and deepest grey indicates no variance.

From this tabulation, it may be deduced that in four examples highlighted in lightest grey, three occur with assessments for *Presentation*, and one for *Task* and *Message*, where it is agreed that the production is "irrelevant" to the task set.

The assessment of *Message* apparently accounts for the different scorings.

Table 6.6

Variance in Evaluation (*Language B, Standard Level*, May 2000)

Candidate	*Assess-ment Criterion*	*Assistant Examiner's comment and scoring*		*Moderator's comment and scoring*	
1	Task, Message	Task completed. Message banal, sometimes almost incomprehensible.	4	Task completed. Message lacking in ideas, sometimes incomprehensible.	3
	Presentation	Inappropriate. Almost no organisation of ideas. No credible paragraph division.	4	Appropriately personal but deficient paragraph division. Correct introduction and conclusion.	3
	Language	Sometimes difficult to comprehend. Overly influenced by translation from English. Very erroneous use of grammar	4	No additional comment	4
2	Task, Message	Task appropriately completed. Ideas well developed and illustrated. Relevant, but genre-selection unjustified	7	No additional comment	7
	Presentation	Essential elements missing for traditional genre presentation. Register consistent. Structuring acceptable.	6	Very good presentation as a debate. Well structured, with appropriate conclusion, though rather abrupt introduction.	8
	Language	Comprehensible for the greater part, sometimes even fluent. Correct but rather limited vocabulary range.	6	No additional comment	6
3	Task, Message	Generally inadequately completed. Ideas undeveloped and repetitive	2	No additional comment	2
	Presentation	Confused. Little evidence of structure. Appropriate register.	3	No additional comment	2
	Language	Very difficult to follow. Inadequate vocabulary. High proportion of grammar and spelling errors.	2	Almost every sentence contains errors. Barely French.	1
4	Task, Message	Comprehensible, but mainly irrelevant to task set. Message repetitive.	3	No additional comment.	3
	Presentation	Apt as diary entry, but not perhaps as 'story'. Clearly structured and developed, but ideas poorly organised.	4	Clear effort to structure presentation.	6
	Language	More or less comprehensible overall, but with significant moments of incoherence due to errors in grammar and vocabulary.	3	No additional comment.	3
5	Task, Message	Largely irrelevant, though easy to follow, detailed and convincing. More depth of argument required	4	Task completion largely irrelevant.	2
	Presentation	Clear and appropriate. Logically structured. Apt register and use of style, but lacking a conclusion.	6	No additional comment.	6
	Language	Comprehensible overall, but with abundant errors in grammar and vocabulary. Little evidence of sophistication.	5	Clumsily expressed. Not always understandable.	4

Table 6.7

Variance in Evaluation (*Language B, Standard Level*, May 2001)

Candidate	*Assessment Criterion*	*Assistant Examiner's comment and scoring*		*Moderator's comment and scoring*	
6	Task, Message	Task well executed. Message easy to follow if occasionally repetitive and superficial. Generally convincing.	7	Agreement signalled. No additional comment.	7
	Presentation	Well organised into paragraphs, with appropriate, if simple linkages established between them.	7	Possibility of more generous allocation of marks within same descriptor level.	8
	Language	Fluent use of language. Some errors repeated. Vocabulary choice occasionally inept. Easy to read.	8	Agreement signalled. No additional comment.	8
7	Task, Message	Task well executed. Varied ideas well developed in authentic and convincing fashion	7	Agreement signalled. No additional comment.	7
	Presentation	Fairly good paragraph organisation. Appropriate register, if rather literary. Apt rhetorical questioning.	8	No additional comment.	7
	Language	Sometimes rather incoherent, but comprehensible overall. Some successful use of sophistication.	4	Agreement signalled. No additional comment.	4
8	Task, Message	Task generally well executed. Relevant but banal. Introduction and conclusion convincing.	6	Inadequate as proposition for action plan, as required. Some irrelevance.	5
	Presentation	Clear and logically structured according to genre. Apt register. Ideas create conviction, rather than use of appropriate transitions.	8	"Ideas rather than use of appropriate transitions create conviction" quoted with comment: "exactly so".	6
	Language	Sometimes difficult to follow, but readable, if overly anglicised. Some successful use of sophistication, marred by occasional incoherence.	4	Agreement signalled. No additional comment.	4
9	Task, Message	Inadequate response to task (letter drafted in past instead of future). Comprehensible but superficial.	3	No respect for genre conventions required (not in diary form).	0
	Presentation	Elementary organisation and cohesive development of ideas. Adequate, if idiosyncratic use of register.	5	*Assistant Examiner*'s judgement too generous.	4
	Language	Often very influenced by English, but fairly easy to follow. Fairly correct use of elementary grammar, with no attempted sophistication. Adequate vocabulary use.	5	Disagreement with statement: "fairly correct use of grammar", otherwise no additional comment.	4
10	Task, Message	Sometimes incoherent and often irrelevant. Unconvincing argument, with elementary, poorly integrated ideas. Little awareness of readership.	3	Agree ment signalled. No additional comment	3
	Presentation	Inconsistent presentation, though often in conformity with formal letter genre requirements. Occasionally too familiar in register.	5	Faults more significant that as assessed by *Assistant Examiner*	4
	Language	Material copied from texts of tasks proposed. Comprehensible but without attempted sophistication. Minimum length barely achieved.	4	No additional comment	3

The *Subject Report* for French *Language B*: November 2000[524]

This relates to detailed observation of **IBO** moderation practice, summarised earlier. Only items directly relevant to the research are described, analysed and discussed[525]. Thus investigated are:

- evident problems of candidate performance, attributable to inconsistencies in **IBO** espoused theory and practice, with analysis of implications for applying assessment criteria and procedures;
- the 'positioning' effects of assessment;
- grounded understandings of authenticity emerging from further analysis of *Chief Examiner Subject Reports*[526].

The November 2000 examination *Report* opens by publishing conversion values for transforming *Paper 1* scores into grades on the **IBO** seven-point scale, as determined at the *Grade Award Meeting* observed. It summarises *Chief Examiners'* general impressions, including comment from teachers, *Internal Assessment Moderators* and *Assistant Examiners*[527], also reporting candidate-number and totalised-score statistics for the session. In comparison with November 1999, overall point-attainment was more concentrated in upper ranges, especially at *Higher Level*, despite claims by two teacher-respondents that the November 2000 examination was "more difficult" than before. The presence of 'bilingual' candidates (for whom the *A2* programme is recommended) was noted as little influencing overall statistical outcomes[528].

Question-by-question description, analysis and comment of *Text-Handling* material follow. Although linkage with *Written Production* is evident (as required by examination-design criteria), task-based relationships between papers are not discussed.

Similar reporting follows for *Paper 2: Higher Level.* Relative task-'popularity' is recorded, measured by total response-numbers for each task of six. Only *Task 6* is specified as 'unpopular'[529]. *Examiners* note candidates' preference for formal-essay presentations, with texts ill-adapted to task-formats, despite the frequent ineptness of such choice.

For the six tasks proposed, *Chief Examiners* are at least implicitly concerned by:

- candidates' knowledge of current affairs, assumed commonly interesting and linked to readings from *Paper 1*[530]. For the **IBO**, general readers would expect 'rational' and 'reasonable' argumentation from satisfactory task-responses;
- candidates' understanding of how to 'convince' typical readerships, as if tasks refer 'authentically' either beyond the assessment world of examinations and examiners to simulated situations, or reflexively to the examination-content itself. In particular, unreasoned, unexemplified, and 'emotional' productions are deemed less 'convincing';
- the reproduction of conventional genre-forms in written task-responses, assuming likely audience or readership expectations[531];
- the acceptance of unexpected, or 'original' interpretations and approaches to set-tasks, departing from established norms, with the proviso that rationales for divergence from presumed readership 'expectations' be justified under *Message*.

In concluding the initial section with *General Recommendations for Future Candidates*, *Chief Examiners* stress the necessity to contextualise *Message, Presentation* and *Language* appropriately within chosen task-simulations. This requires candidates to recognise and imagine 'real' worlds beyond the examination's specific representations. Teachers and students are therefore exhorted to heed not only likely expectations, but also specific audience or readership needs and interests, set in appropriately 'authentic' environments and addressed through appropriate linguistic and cultural forms.

Simultaneously, candidates are encouraged to choose tasks for which they are linguistically well-prepared, rather than prefer those that otherwise focus attention or arouse interest. The recommendation directly contrasts with Van Lier's approach to identifying authenticity in communicative language-production[532]. Examinees are explicitly warned however, **not** to link pre-learned idioms through stylistic exercises, assumed 'appropriate' for examination purposes. This lacks "authenticity", by restricting the display of "competent fluency"[533]. The dilemma prefigured goes to the heart of issues raised by the research questions.

In reporting *Standard Level* candidate-performance, a similar format is used, but with *Chief Examiners'* observations related in greater detail.

Teacher-responses are summarised, with sampling from the paper-design and content surveys by questionnaire that accompany each examination. They indicate that in November 2000, *Paper 1* was considered approximately equivalent in 'difficulty' to the previous year's paper. Clarity of rubrics, question-styles and presentation were "satisfactory"[534]. For *Paper 2*, the *Examiners* reported 20 returns[535], with

19 recording tasks and proposed content as “appropriate to the level of knowledge and experience of the candidates”. One teacher replied that the session had been “slightly more difficult”. All stated a “general[..... satisfaction] with the variety of themes and tasks required”[536].

However, the *Examiners* note *Paper 2* as the “most difficult” component, requiring “training” in technique throughout the two-year programme. Preparation should involve careful analysis of tasks, “judicious” choice of response, knowledge of appropriate language, genre and presentation forms expected by ‘typical’ readerships, with good content planning, prior to composing clearly-legible and uncorrected final drafts[537]. Given one case of supervision error at this session (candidates from a single centre having completed work on *Paper 2* with *Paper 1* texts at hand)[538], the *Chief Examiners* remind *Report* readers of the need to respect examination regulations.

Following further generalities, the conversion scale is published for transforming moderated, *Assistant Examiner* ‘raw’ scores to **IBO** grades, by equivalences determined at the December 2000 *Grade Award Meeting*. In order of popularity, candidate choices are summarised, with tasks Nos. 4 and 6 noted as least popular, and Nos. 1 and 2 (respectively linked to *Paper 1* texts C and A) as most popular. Certain candidates poorly adapt their messages to the specific tasks proposed, simply reproducing ideas from *Paper 1* readings[539].

In detailed comment concerning the six *Standard Level* tasks for November 2000, the following preoccupy the *Chief Examiners*, and supplement their prior remarks:

- inadequate text contextualisations not only within definable social situations, but also within 'likely' psychological relationships between 'self' as author, and 'other' as reader of responses[540];
- insufficiently close adaptation of ideas to specific task-forms[541];
- poor adaptation of language and 'tone' to styles characteristic of task-domains or 'genres'[542];
- deficient respect for social convention when simulating communication within defined relationships[543];
- weak text-structuring for producing intellectual 'argument', likely to 'convince' simulated readerships.

Within *Conclusions and Recommendations*, longitudinal performance measurements from previous examination sessions are reported. Issues of language, assumed appropriateness of subject-choice and careful attention to detail in presentation, are emphasised and made explicit. In examination preparation, teachers and students are thus exhorted to:

- consider simulated audience and readership expectations as constraints on absolute freedom of choice in expressing 'self', as if recipients are indeed also response-assessors. Satisfactory productions will 'convince', a quality *Examiners* frequently stress in referring to the *Assessment Criteria*[544];
- draft and edit responses, clearly identifying key ideas and language and producing texts that may be read as maximally "authentic", through the avoidance of repetition and 'wordiness'[545];
- respect elementary French grammatical convention[546];
- strive for legibility in handwriting[547];

- respect word-count rubrics for minimum length in assessed *Written Productions*[548].

The *Report* concludes by publishing tables for converting the aggregated, *Internal Assessment* component scores into corresponding grades on the **IBO**'s seven-point scale, with comments on this aspect of teacher and candidate performance.

In preliminary remarks, the *Chief Examiners* record that moderated oral performances are generally high in standard. However, they lament the paucity of candidates choosing francophone cultural themes, or comparing a francophone culture with their own[549]. Implicitly, they confirm the view that free topic-choice, a prerequisite for authentic communication according to the research criteria, positively influences outcomes in oral presentations and ensuing discussion. The *Chief Examiners* identify qualities such as:

- clear and precise exemplification of current affairs issues through apposite research[550];
- 'good' development of personal opinion through critical reasoning[551];
- committed motivation to communicate with others, expressed through lively interest in debating polemical subjects[552];
- stimulating interaction in developing communicative dialectic between 'self' and 'other', allowing teacher-assessors to stimulate and challenge orally-expressed views[553];
- demonstrable spontaneity in language-production in interchanges freed from negative constraints and nervousness induced by 'positioning' within the assessment context[554];

- 'authenticity' in debate, subsuming the above and emphasised by seamless progression through time, in communication of unvarying linguistic standard that typifies much of the performance[555].

In identifying such qualities, the *Examiners'* specifications harmonise with Van Lier's criteria for authentic language use, as further developed in research experimentations and noted on each occasion in the listing above.

Remarking on the difficulties of **IBO** criterion-based assessment, the *Examiners* note that oral productions require sufficiently extensive presentations and discussions to allow application of all the criteria. However, performance should neither be so lengthy as to diffuse the focus of the chosen theme, nor so generalised that extended discussion and appropriate probing through debate become difficult. Teacher-assessors, as facilitators of successful performance, are reminded that interventions should prevent memorised recitation of pre-learned material. This is deemed detrimental to candidates' creation of "convincingness and authenticity"[556]. The remarks further exhort teachers to vary questioning for homogenous candidate groups, thus reducing scope for 'question-spotting' and pre-planned response. The development of genuinely spontaneous initiative and interchange is most desirable[557].

Potentially significant in assessing authentic language use, teacher-assessors are recommended to prevent interruption of audio-tapings throughout recording and to restrict the stages of oral presentation, presentation-discussion and general conversation to maxima of 2/3 minutes, 3/4 minutes and again 3/4 minutes, respectively[558]. It is plausible to infer that reiterations point to a relatively common occurrence of 'irregularities' in practice. Indeed, evidence supporting such inference is presently described.

In recommending pedagogical approaches for preparing assessed performance, the *Chief Examiners* suggest:

- avoiding "banal" presentation topics, since these limit the "value of the discussion";
- testing candidate '"aptitude" to defend personal opinion and respond appropriately to counter opinions put by the teacher-assessor, especially at *Higher Level*;
- focusing attention on "correct pronunciation";
- avoiding interruptions and error-corrections during oral presentations[559].

These concerns demonstrate teacher-assessor 'power' in positioning *Internal Assessment* candidates, constraining their choice of language, message and presentation, and determining levels of interaction. Not only are examinees to be 'guided' towards 'appropriacy', but linguistic norms are judiciously to be considered. Candidates should respect assessment contexts and perform within criterion constraints, as well as conform to cultural and linguistic norms adopted by teacher-assessors. Authentic language use may thus be facilitated, satisfying Van Lier's conceptualisations of *Creator Authenticity and the notion of 'other', Authenticity of Context* and of *Interaction*[560]. Nonetheless, teacher-assessors may evidently compromise the latter through overly influencing candidate-choice and freedom to construct interactive communication. Authoritative insistence on reproducing modelled, 'successful' performance may become favoured method for preparing assessment-productions, the implications of which are analysed in the following section.

Oral Language-Production for the May 2001 / 2002 Examining Sessions

In furthering grounded research for conceptualising authenticity and understanding associated **IBO** theory in practice, as well as in further investigation of the constraints of 'positioning' by **IBO** programmes, empirical evidence for oral performances were scrutinised in detail.

Under duties as *French Language B Internal Assessment Moderator*, 55 oral productions from 12 examining centres in Canada, The Netherlands and the United States for May 2001, and 52 productions from 10 US centres for May 2002, were analysed and are reported. These samples, selected by teacher-assessors or candidates (either individually or in joint negotiation), were recorded at examining centres, being mostly produced under programme-rubrics as obligatory components for validating reliable *Internal Assessments.*

With performance evaluated at a maximum of 30% of total marks available[561], they are reassessed and moderated as individual, candidate-presentations of candidate-chosen topics, drawn from one of three, broadly-specified domains[562]: *'Exploration of Change', 'Exploration of Groups',* and *'Exploration of the World of Leisure'*[563]. In practice, wide variation is evident, with researched samples providing evidence of a large range of themes and performance-durations as low as eight or nine minutes, and as high as twenty minutes and more, rather than approximating the twelve minutes recommended.

In programme-descriptions[564], authentic production and interchange are linked by implicit criteria to language-selection, content, structuring and presentation as task-based response, theorised in previous chapters. As with *Written Production*, they may be analysed, tabulated, and thus made explicit. Relationships with Van Lier's criteria for authenticity are postulated, as shown below and overleaf:

Table 6.8

Principal Characteristics of the Assessment Criteria[565]

Criterion A: Task/Message

General Definition of Assessed Qualities:

The speaker's effectiveness in completing prescribed tasks and communicating appropriate messages.

Category Assessed	***Concepts Assessed***	***Criteria for Authenticity***
Overall Competence	Interest in content	All
Task Completion	Success attained	*Finder* and *User Authenticity, Authenticity of Context, Purpose* and *Interaction*
Message Intelligibility	Clarity and Plausibility	*Creator Authenticity*
Quality of Ideas	Quality of development and exemplification of arguments: relevance, interest and convincingness.	All

Table 6.9

Criterion B: Interaction

General Definition of Assessed Qualities:

The speaker's effectiveness in maintaining flow in discussion.

Category Assessed	***Concepts Assessed***	***Criteria for Authenticity***
Overall Presentation	Liveliness in interaction	All
Interaction with Teacher-Assessor	Willingness and preparedness to participate in dialogue	All, except *Finder* and *User Authenticity*
Cohesion and Ease of Flow	Ease and coherence in exchanging ideas	*Creator* and *User Authenticity, Authenticity of Interaction*, and especially, *Autotelic Authenticity*
Reactions	Comprehension of spoken language and appropriacy of response	All. except *Creator Authenticity as the realisation of 'self'*

Table 6.10

Criterion C: Language

General Definition of Qualities to be Assessed:

The accuracy, appropriacy and fluency of oral production.

Category Assessed	***Concepts Assessed***	***Criteria for Authenticity***
Overall Impression	Ease in language use	All
Vocabulary Choice and Register	Adequacy and variety in the choice of vocabulary and idiom	*Finder* and *User Authenticity, Authenticity of Context, Purpose* and *Interaction*
Accuracy	Precision and variety of grammatical structuring	*Finder* and *User Authenticity, Authenticity of Context* and *Interaction*
Pronunciation and Intonation	Contribution of enunciation to the flow of communication	*Creator* and *User Authenticity, Authenticity of Context, Purpose* and *Interaction*

In comparison with **IBO** criteria, Van Lier's conceptualisations of authenticity are more holistically categorised, with most applying to assessment in any domain. They refocus perspectives through emphasising measurement of sociolinguistic aspects of communication as situated interchange between two or more participants. For triangulating research, they eschew more traditional, positivistic, psychometric, structural and purely linguistic, performance-measurements of language-knowledge, communicative skill, and organisational competence, typified by *Language* criteria.

Following analysis of 107 presentations, few were found to relate explicitly and straightforwardly to themes, broadly specified, although implicit relevance is often easy to verify[566]. Through prescribing domains, the **IBO** 'suggests' underpinnings and directions for guiding teachers and candidates towards satisfactory performance.

Illustrated by the topic data-base listed in **Appendix 4**, categorisations by theme appear tenuous, with such varied candidate-choice that **IBO** prescriptions effectively form three, all-embracing sets. No single, given topic (even those judged too simplistic, or too complex, for the programme-level presupposed) could easily be excluded[567]. For examinees in social and linguistic interaction between 'self' and 'other', provisions of wide-ranging options bring authenticity into play. With many topics dedicated to personal themes selected from no prescribed French curriculum, diversity evidently permits candidates to commit themselves to presenting and discussing subjects of individual interest. Indeed, in no recorded case did candidates produce language immeasurable under qualitative assessment by either **IBO** or experimental criteria. Descriptions of minimal attainment were almost completely absent[568], with no patterns of preference apparent. Hence, the availability of genuine choice seems real. Occasions for exercising authentic autonomy are offered and integrated within the **IBO** design for *Internal Assessment*. *Creator Authenticity* comes to the fore.

However, greater investigation of evaluation procedures is required, with further moderation of non-examinable, oral performances normally excluded from the agenda of *Grade Award Meetings*[569]. In *Figure 6.4*, problematic assessments were identified as points falling outside the broad diagonal described by the scatter graph. Closer inspection of pertinent cases reveals clear anomalies.

In *Subject Reports*, *Chief Examiners* record a general trend of rising standards, in part tentatively ascribable to 'washback', through growing candidate, and assumedly teacher-familiarisation with examination and assessment expectations and 'styles', as well as with techniques favouring success, if not through improved pedagogy and learning *per se*. 'Washback', as the facilitation of 'teaching to examinations' through assessment-design, evidently affects the authentic use of language as interactive communication between 'self' and 'other': that is, the design of **IBO** assessment criteria and the exhortations of *Subject Reports* may favour particular teaching styles and learning techniques that hinder authentic expression. Examples from highly anomalous cases appear to illustrate the feature.

By assimilating techniques through 'washback', together with the knowledge and skills they promote, candidates may respond more successfully to subjects and tasks proposed under any given rubric. The necessary integration of 'self' with individualised purpose and perspective, in linguistic productions recognised as 'appropriate' to set assessment-tasks and representing meaningful attempts to communicate through the authentic use of language, may be compromised or even 'faked'. This exploits dichotomies between examinations as opportunities for situated communication, albeit in awareness of future assessment, and task-based response as simulations of communication in extra-examination contexts[570].

The micro-sociolinguistic phenomenon of 'interlanguage', familiarly created in interchanges between teacher-interlocutors and candidates who over time, 'know' each other increasingly 'well', poses additional problems. For certain cases and across certain combinations of languages (especially those with close structural and lexical relationships, such as French, Spanish and Italian, or Dutch and German), the phenomenon may have significant implications for assessment and evaluation. Candidates may reproduce various dialectal, or Creole language-forms, 'authenticised' by intimate reference to personal usage by 'self', though potentially impeding communication with outside parties[571]. In such situations, implications in assessing authenticity and equity, should be considered. They increase in importance when other languages, commonly known to candidates and interlocutors, serve as aids, or act as barriers to comprehension and communication. As such, the question of language 'distance' (or degrees of commonality between any given pair of languages)[572] needs further research.

Indeed, analysis of moderation practice emphasises the effects of 'interlanguage' use on authenticity in *Internal Assessment*. Authentic expression may represent recourse to personal strategies for communicating messages between two individuals who have come to 'understand' each other. Candidates and interlocutors may validate these in common. For any jointly communicative, dialectical construction of linguistic interaction, interlanguage characterises intersubjective interchanges between 'self' and 'other'. It forms a cornerstone of any ontology and epistemology for phenomenological conceptualisations of authenticity, especially those underlying the research.

In contexts where social power-relations are unequally distributed and outcomes have 'high-stakes' value, the phenomena, it is surmised, result partly from 'micro-acculturation' and familiarity gained through length of acquaintance, and partly from pedagogical concerns amongst teacher-assessors in 'authoritative' roles as 'facilitators'. Evidently a subject's preliminary choice to enter into communicative action, interpretatively constructing meaning and understanding, is however, fundamental.

Reliable assessment-performances, recorded as relevant interlocution, whether oral or written, are produced through subjective volitions 'uncovering' or 'creating' such meaning and extending into continued interchange. The resultant communication is 'private', as intended by participants, but not necessarily either wholly or partially comprehensible to third-party audiences, ignorant as they are of the particular status of the actors in communication. For the research, audiences are defined by *Internal Assessment Moderators*, personally and professionally 'unknown' to interlocutors.

On recalling that individual productions are assessed, moderated and evaluated, even though intimately integrated with, and thus dependent upon the interview performance of teacher-assessors as occasional initiators and full co-respondents, the problems become more evident. Indeed, *Chief Examiner* comment and recommendation constantly stresses mindfulness of candidate and teacher interdependencies and the latter's assessment responsibilities[573]. In cases of maximal performance according to Van Lier's criteria for *Authenticity of Interaction*, ideal communicative integration between two individuals should demonstrate equality in social relations, and in effect be 'total', displaying 'symmetry' in participation rights and duties, and creating evidence of autonomy for 'self' in intentional interactions with 'others', recognised as equally autonomous[574].

For measurements of authentic language use, the implications of **IBO** *Internal Assessment* criteria and procedures, as well as of equity in anomalous cases, are illustrated most clearly by exceptions, identified in *Figure 6.4*.

Usually, anomalies derive from form, content and exchange in oral production, when teacher-interlocutors intervene as *Internal Assessors* seemingly concerned to ensure success for their respective candidates, yet following apparently behaviouristic modes in structuring activities. A single assessment centre provides the clearest examples from May 2002[575]. Here, it seems probable that candidates were not discouraged from reading pre-prepared texts as speeches. The ensuing interviews were dominated by closed questioning from the assessor-interlocutor, 'authentically' eliciting many monosyllabic, 'yes' or 'no' answers, yet providing insufficient evidence of interactive communication to allow full application of the assessment criteria. Otherwise, teacher-interlocutor statements were so lengthy as to occupy the larger portion of the time available for recording productions. It was also apparent that language 'correctness' and elaboration varied so markedly between presentation and interview, as to question the respect of rubrics, defined in the *Guide to the Programme*[576].

Moderation evidence from this single centre illustrated candidate-reliance on pre-prepared, routinely-memorised statements (albeit not always appropriately so), often appearing as recitals (occasionally inexpert) of written texts, sometimes inserted into conversations with little respect for logical cohesion, perceptible by the listener.

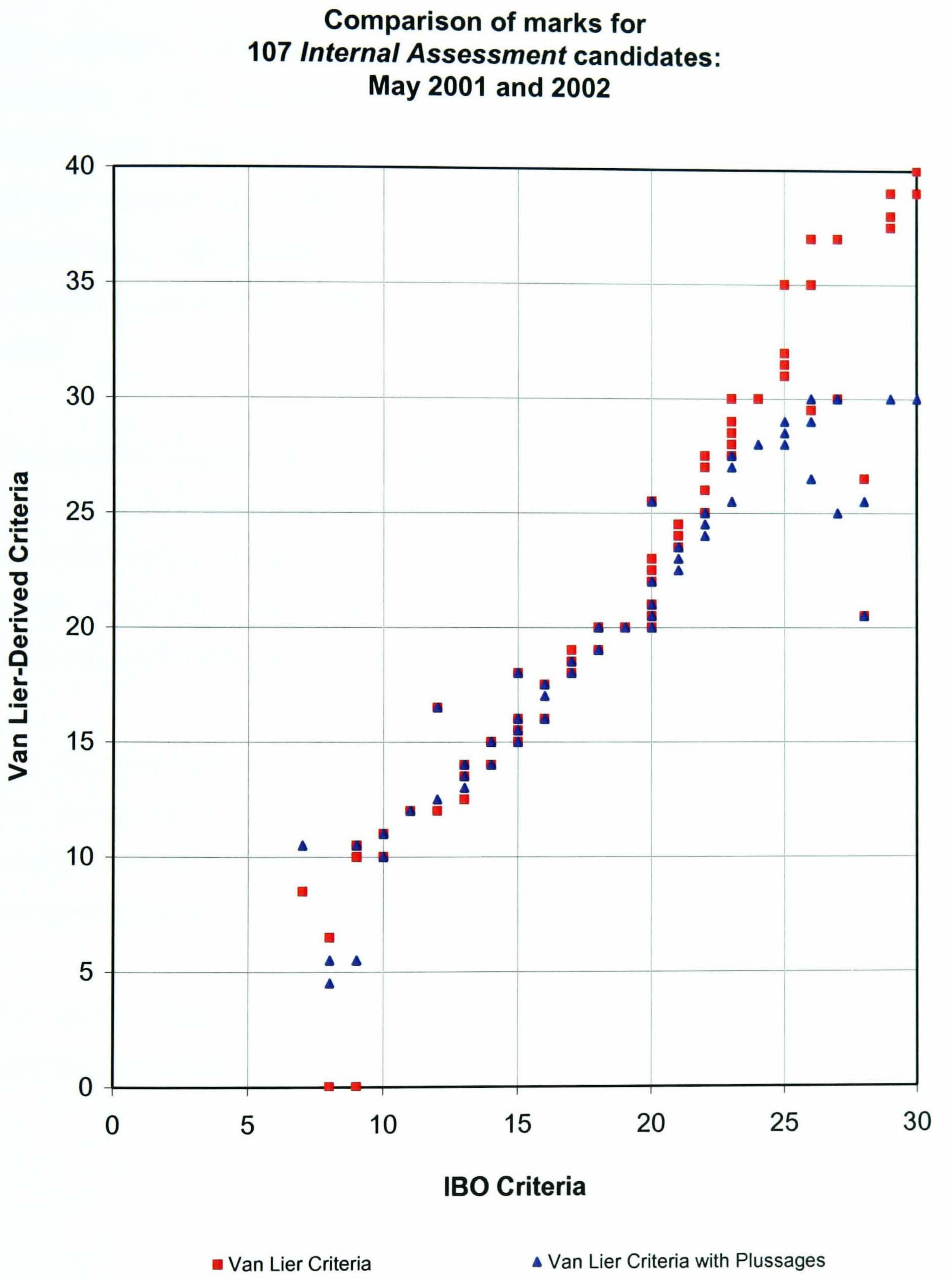
Authentic communication is thus minimally achieved, with opportunities for relatively unconstrained self-expression in interactive interchange with 'other', reduced by the dominance of *Internal Assessors* as teachers. Candidate failure to provide sufficient, orally-produced language may severely impede performance to the detriment of resultant assessment.

The phenomenon may relate to deeper issues for conceptualising authenticity, emerging from the experience of *Internal Assessment* moderation. Discussion of data requiring more thorough analysis, now follows.

The scores allocated by the *Moderator* in completing **IBO** duties, and by the researcher in experimental re-assessments for triangulation, were recalculated for graphical representation. Variable teacher-assessments lying outside the scope of the exercise were excluded, since the rationale assumes stability under a single assessor's criterion-interpretations, even though the statistical validity of results is limited for generalising.

In two cases, data for *Internal Assessor* scores were missing, explaining those points valued at zero, with totals represented in *Figure 6.5*, overleaf.

Figure 6.5


Comparison of marks for
107 *Internal Assessment* candidates:
May 2001 and 2002
Van Lier-Derived Criteria
0
5
10
15
20
25
30
35
40
IBO Criteria
0
5
10
15
20
25
30
■ Van Lier Criteria
▲ Van Lier Criteria with Plussages

Preliminary inspection of these quantitative representations is surprising for being unexpected.

For example, although for any given individual presentation and interview, score variations between *Internal Assessor* and *Internal Assessment Moderator* allocations could be large, and indeed for that reason subject to further moderation by the **IBO**[577], the general distribution, representing 107 assessments from two sessions[578], produce a discernible pattern when arranged in ascending, numerical order.

The pattern demonstrates consensus, in that not only do 107 gradings form a roughly diagonal line indicating close score-correlations across the entire range, under either assessment system, but also that few are clearly aberrant. These may be isolated from 'typical' cases for more detailed, qualitative scrutiny, as presented earlier.

The Van Lier-derived system, with plussages aggregated to a maximum forty points, rather than a normal thirty, differentiates performance more clearly, and with fewer outliers at higher levels of attainment, whilst demonstrating close correlations in remaining cases[579].

The same data has been numerically analysed and tabulated in summary overleaf:

***Grade Differences*:**

Plus 3 or more:	0%
Plus 2:	6%
Plus 1:	38%
None:	*56%*

For which:

Minus 3 or more:	0%
Minus 2:	1%
Minus 1:	5%

****

Grade Decrease of 3 or more:	0%
Grade Decrease of 2:	1%
Grade Decrease of 1:	5%
No Change:	*56%*
Grade Increase of 1:	33%
Grade increase of 2:	5%
Grade Increase of 3 or more:	0%

Raw Score Differences:

Negative:	**Positive**:
- 7.5 : 1%	0.5 : 7%
- 3.5: 2%	1.0 : 24%
- 2.5: 2%	1.5 : 7%
- 2: 1%	2.0 : 12%
	2.5 : 4%
	3.0 : 13%
	3.5 : 3%
	4.0 : 4%
	4.5 : 1%
	5.0 : 3%
	5.5 : 3%

None:

13%

AVERAGES:

Average raw score on **IBO** system:	18.51 / 30
Average raw score on Van Lier system:	19.98 / 30
Average Grade with **IBO**:	4.63 / 7
Average Grade with Van Lier:	4.99 / 7
Average difference in raw score, Van Lier to **IBO**:	+ 1.52 / 30
Average difference in Grade, Van Lier to **IBO**:	+ 0.36 / 30

Written Language-production for the May 2000/2001/2002 Examining Sessions

In May 2001, as further response to the same research questions, the assessment of 154 scripts from five different centres in Canada, Switzerland and the United States, was completed according to **IBO** assessment criteria and procedures. Detailed reports representing the work of 81 candidates from three differing centres (in Switzerland and the United States) were also composed according to the rubrics for *Individual School Reports*, by the researcher-*Assistant Examiner.*

In May 2002, similar data-collection covered a total allocation of 151 scripts produced for *French Language B, Written Production* from nine different centres located in Bahrein, Canada, Kuwait, Oman, Saudi Arabia, Syria and the United States. Detailed reports representing the work of 27 candidates from two different centres (in Canada and the United States) were composed according to **IBO** rubrics in completion of examining duties.

From professional experience, 'interesting' productions were noted and scatter graphs illustrating variance between **IBO** and experimental, Van Lier-derived scorings were created.

Once again, in patterns illustrating broad correlation with oral re-assessment findings, points outlying the 45-degree diagonal around which most cases were clustered, indicated 'exceptional' variance. As with *Internal Assessment*, these performances were deemed in some fashion 'aberrant'. On further scrutiny, they were found to correspond with those noted from initial assessments under examining duties.

In evaluating authenticity in writing however, the research strategy was modified, potentially significantly. In only a few cases was contemporaneous assessment under Van Lier criteria completed, since little time was available to meet tight, **IBCA** deadlines. In the problem notwithstanding, there lay an advantage.

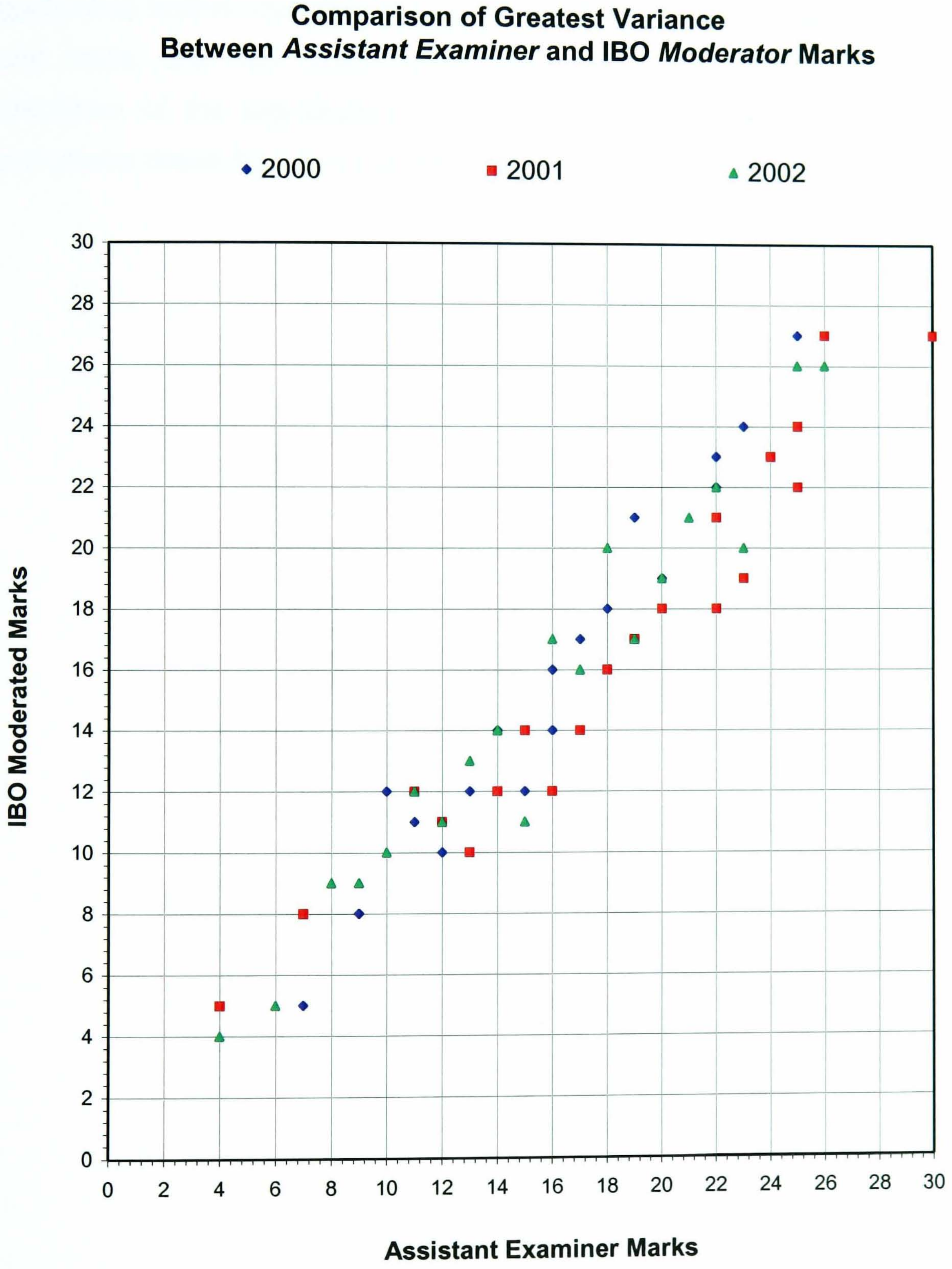
By introducing a time-lapse into processing scripts, new and useful perspectives on the stability of assessor understandings and consistency could be obtained. That is, a perspective was created for viewing **IBO**-sampled, *Written Production* assessments[580] not only over time, but also in different research triangulations and analyses of emergent understandings[581].

Furthermore, despite the development of Van Lier criteria within a single model in two distinct cycles, it was possible to review pilot-research data, by reworking *Written Production* assessments from the May 2000 examination session. The results were represented graphically, as before.

With data for simultaneously-completed assessments in May 2000 and 2002, and after a time-lapse in May 2001, comparisons demonstrate assessment 'stability' for a single *Assistant Examiner*, over a three-year period. Further **IBCA** data, measuring reliability through comparing sampled *Assistant Examiner, Team Leader* and *Moderator* assessments, confirm respect for **IBO** standards, as noted previously.

Thus, more reliable results are represented in *Figure 6.6*, overleaf.

Figure 6.6


Comparison of Greatest Variance
Between *Assistant Examiner* and IBO *Moderator* Marks
◆ 2000
■ 2001
▲ 2002
IBO Moderated Marks
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Assistant Examiner Marks

Additionally, the biannual *Subject Reports* are important for *Chief Examiner* comment on **IBO** moderation and evaluation processes. Indeed, *General Grade Descriptors,* relating final evaluations through aggregating scores and grades to 'appropriate' points on the **IBO** seven-point scale, are only made widely available in these publications. Discussion of the significance of this process is integrated with the conclusions drawn from the evidence and presented in Chapter 7.

PART IV

CONCLUSIONS

CHAPTER SEVEN

AUTHENTIC DESIGNS
FOR MODERATED ASSESSMENT AND EVALUATION

Preliminary Conclusions

From the evidence, rubrics, tasks and criteria for assessing and evaluating **IBO** *Group 2: Language B* productions appear coherently designed. Whether formally assessed or not, scope for authentic communication is broad. For *Written Production,* familiar, socio-culturally appropriate interactions are explicitly contextualised and offered in option. In *Internal Assessment*, framings are partly chosen by candidates themselves. Through integrating all discrete language skills within intersubjective, interactive and knowledgeable usage, traditional, positivistic, norm-referenced and non-interactive approaches to testing and assessment are eschewed. In fulfilling **IBO** requirements, candidates may personally, imaginatively and convincingly link 'self' to 'other', through chosen topic or task response, with greater or lesser degrees of recognition. The design for production largely, though not wholly, satisfies major criteria for linguistic authenticity.

Listening and speaking are combined in *Internal Assessment*, albeit in varying amounts. Oral performance partly depends on aural competence, yet is privileged through quantitative weighting, even if hypothetical, non-listening speakers may be assessed above a minimum zero. Accordingly, for aggregations of componential scores, listening skill appears under-recognised, under-assessed and under-represented.

Formal *Written Production* combines reading and writing, though strongly emphasising the latter in evaluation, with reading separately assessed through *Text Handling*. Exceptions occur with tasks thematically related to *Paper 1* readings.

Here, it may be recalled that many traditional schemes assess skills discretely, often behaviouristically, through structural, psycholinguistic and psychometric approaches. For 'objectivity', choice by self is largely restricted and favoured through uniform equalisation of tasks, with scripted prompts tightly constraining expression. Reproducibility in controlling for reliability is of greater importance than purest construct validity, with productions required to match predefined normative utterances as 'model' responses[582].

Construct Validity in IBO Task-Designs

Considering the evidence, few *Diploma Programme* candidates completely misunderstand designers' task intentions, thereby scoring poorly in assessments under either system researched. Nor do many apparently misconstrue task constructs to their own detriment. The conclusion is negative, deriving from a relative absence of data, as sampled. Indeed, 'gatekeeping', by which teachers or schools select 'appropriate' candidates for examination entry and assessment, may account for the finding. Explanations of the phenomenon lie beyond the research bounds and remain uninvestigated. The key issues concern weightings of performance-values per criterion, and their effects on construct validity for quantified evaluation, whether authentic or not.

In researched assessments, overall rater-variance by aggregations fell mostly within accepted norms with little unreliability, validating comparisons. Quantifications appear soundly founded, permitting measurement of situated, task-based, authentic performance, whether in *Internal Assessment* or *Written Production*. However, occasionally, potentially significant variance was found in results per criterion, questioning the overall construct validity of componential aggregation, *per se*.

Similar observations hold true of 'washback' effects, threatening the authenticity of valid constructs through promoting imitation of non-authentic, replicable and decontextualised models of high-performance. (Such is possible in highly-constrained, highly-predictable and more strongly positivistic systems, such as *GCSE*[583]). For the **IBO**, teaching and learning 'to the test' are made possible and enhanced through organisational commitments to publicise final-design, criterion and procedural data in various formats, not least for familiarising teachers with such descriptions through recommended, workshop training, based on collective study of exemplars of past performance. Prescriptions of minimum, model language are avoided. The policy ensures transparency, advocating authenticity in all aspects of production. It promotes common understandings of task-purposes, standardisation, assessment and evaluation amongst **IBO** personnel, teachers and students alike.

However, 'washback', promoting 'high-stakes' performance, is associated with standardisation, since this intentionally facilitates predictable uniformity of task as its goal. Growing teacher and candidate familiarity with fixed designs and language 'levels' may account for certain empirical findings, though such influence on teaching, learning and curriculum-content selection, has not been extensively researched. Indeed, with *Internal Assessment*, formative pedagogy and criterion-based evaluation are explicitly recommended for stimulating regular and continual improvements in student productions. Amongst the samples researched, this partly explains relatively low, and apparently decreasing incidences of extremely low-scoring attainment.

Construct validity relating task-requirements to candidate-performance is mostly high, as underlined by experimental re-assessment. Any programme promoting authentic expression through interactive communication between situated subjects would require as much. The results obtained from manipulations employing Van Lier-derived criteria appear closely and reliably correlated to those recorded for the **IBO**, (given limitations of method through recourse to a single rater). The conclusion is constrained however, by primary, *a priori* choices of research scope and bounds.

Reliability in Assessment and Evaluation

For the **IBO**, reliability is established through standardising all assessment-tasks, and repeatedly applying securely criterion-referenced, qualitative moderations in interpretative triangulations by assessors. It seeks validity through consensual verdicts assessing intersubjective relations between speaker and listener, writer and reader. Based almost exclusively in re-assessment under discrete, though interlinking criteria, broadly coherent in specification and consonant with theories of linguistic authenticity, the results are published as transformations of quantitatively-aggregated evaluations into numerical, final grades. Productions are not compared with prescriptive 'model'-answer references (were such possible), devised and circulated by the organisation, in 'objectivised' 'marking' schemes.

Longitudinal adjustments, effectively norm-referencing standardisations of knowledge, ability and performance for sets of candidates over time, are minor in effect. Whilst evidently influencing *Language* assessment in oral and written modes, they do not overly constrain the design, implementation and outcomes of individual, task-based performances. For mutually-independent programme groupings, moderation procedure determines and justifies linguistic 'standards' based on tasks proposed, seeking equable coherence and consistency within each programme. Such 'standards' are not immutably fixed within predefined gradations of level.

Despite some arbitrariness in specifying groups as *Ab Initio, Language B* and *Language A2*, and in defining *Standard* and *Higher Levels* for the latter, *French Language B* task-design and its standardisation are plausible. Maximum attainment, matching the 'highest' qualitative descriptions, is both possible and regularly achieved. Whether under **IBO**, or experimental criteria, research assessments amply demonstrate the finding. The professional experience of task-designers and standardisers effectively ensures that tasks and rubrics are neither too facile, nor too demanding for 'typical' candidates, when assessed under relevant criteria.

Interpretative Intercommunication

Responsible examiners exercise ultimate powers of decision in design and standardisation. In certain cases, ambiguities grant significant leeway in interpreting and implementing discrete, programme requirements. Nevertheless, the various arrangements appear well understood by the **IBO**'s clientele, supported as it is by intensive, recommended and regularly-held, training workshops for programme familiarisation. The high incidence of performances respecting both rubrics and tasks, and permitting relatively unproblematic assessment according to specified designs and procedures, points to broad acceptance of the system and its content by 'stakeholders' worldwide: candidates, teachers, designers, standardisers, examiners, moderators and those university-level supervisors ultimately validating the whole in its international context.

Ideally and often practically too, the programme encourages candidates positively to respond with social creations of communication in a given language, participating at appropriate levels of competence. In limited experimental range, the results provide stable productions, adequate for identifying, measuring and differentiating the quality of key, discrete elements of authentic expression. Whether orally or in writing, almost all record greater than minimal attainment under each category. Exceptions relate largely to individual teacher-facilitators and assessors failing correctly to observe *Internal Assessment* rubrics and thereby constraining appropriate candidate participation. The overall design allows most however, to display clear linguistic evidence for creating and maintaining *Curricular, Pragmatic* and *Personal Authenticity* in all their sub-divisions. Successful performance interpretatively requires situated, intersubjective interaction between 'self' and 'others' as listeners, interlocutors and readers, (if not always as writers, given a single channel of communication, lacking interactivity in 'feedback' through textual replies under formal examination). Language is linked to social contexts, loosely delimited by assessment-settings, yet stimulating authentic expression.

Within research bounds, the findings indicate positive candidate interpretation of free choice of task, genre and content. Under *Internal Assessment,* the presentation of the interests of 'self' is open, in effectively unlimited range. This remains true, albeit only partially, in more constrained cases of *Written Production*, with six tasks provided in option for situated, socio-culturally appropriate response. Attainment at the highest levels of description requires language-production adequately recording relevant subjective qualities in their creators. In *Standard Level Internal Assessment,* these are revealed by candidates who commit individual 'selves' to communication "with ease and fluency", with "some feeling for the language" and "a degree of sophistication" whereby the "listener is interested and drawn into the flow". In *Written Production,* the highest-valued language will illustrate ideas that are "original and/or convincing". The qualities rewarded are emphasised at *Higher Level,* with highest general descriptions available for those "communicat[ing] with an air of authenticity"[584], all according to criterion-based, assessor interpretations.

Positivistic Concerns in Assessment and Authentic Criterion-Referencing

Whilst "competence" in producing "accurate language", and "variety of vocabulary and idiom" form criteria relating to structural 'standardisations' of language, as uniquely valued by 'others'[585] and indicating implicitly-favoured, psycholinguistic and non-interpretative approaches to evaluating second-language acquisition through norm-referencing, their quantitative value-weighting in performance-scores is minor. Major, criterion-referenced weighting is devoted to culturally-based and sociolinguistic categories of appropriate presentation, allied with relevant task-completion. All categories value 'existential' concerns for self, expressed through willing participation in linguistic communication, and originality in individual perspective. Personal awareness and autonomy, realised through choosing tasks, may be fully integrated within individual responses, respecting programme-design and assessment practice. Evidence for highest attainment must match interpretative descriptions of the most demanding criteria, requiring demonstrations of 'personality', 'imagination' and ability to 'convince' audiences or readerships, within clearly-specified communicative roles, complementing those of assessor, examiner or moderator.

From research data, **IBO** assessment-tasks emerge as inclusive, non-discriminatory, and explicitly embedded within relevant historical, cultural and linguistic contexts. No contrary evidence was uncovered. The designs appear to stimulate appropriate response, with genuine, personalised, communicative value, almost completely according with stated programme aims and objectives. They allow for clear differentiation in measuring responses[586].

To most candidates researched, the rubrics for producing assessable language, whether in speaking or writing, appear clear and unambiguous, with little evidence to the contrary. No significant 'hidden curriculum' of assessment norms and values emerges from research. Ultimately, successful task-completion appears both feasible within the constraints, and accessible for all candidates prepared according to programme requirements. Hence, **IBO** claims to high success rates are plausible.

Conclusions and *Internal Assessment*

Internal Assessment arrangements evidently offer greater opportunities for candidate control of 'self'-expression, with lesser degrees of 'positioning' through task-setting and format. They afford greater degrees of freedom for individual choice of content and presentation style, allowing relatively unconstrained expressions of *Curricular, Pragmatic* and *Personal Authenticity* over the extended time-span of assessment. Intersubjective interactions should be (and mostly are) largely 'unrehearsed', 'meaningful' and not explicitly 'artificial'[587]. They are neither scripted, nor prompted as in positivistic and behaviouristic schemes, typically measuring appropriacy and quality by matching responses to restricted, model predictions[588]. Through the transparent positioning of assessors as interested interlocutors and facilitators, **IBO** rubrics create greater scope for genuinely communicative, interactive, intersubjective and hence authentic language use between two or more speakers and listeners[589].

The evidence demonstrates that successful *Internal Assessment* performances depend critically upon the understandings and efforts of teachers as *Assessors.* In representing organisational authority in workplaces, teachers must encourage full knowledge of, and respect for assessment rubrics, respect these themselves, thus guiding candidates to success. Many anomalous performances are attributable to deficiencies in this context. Teacher-assessors must, for example, discourage reading aloud of continuously-sequenced, pre-prepared texts as oral presentations, since this misrepresents assessment aims and objectives for listening and speaking, unduly restricting full application of criterion descriptors to any language-sample thus produced. High performance is significantly more difficult to attain. The point is repeatedly stressed in *Subject Reports*, suggesting a recurrent and fundamental problem.

As interlocutors, teacher-assessors must responsively attend to the ideational content of presentations, encouraging informed discussion of points of interest, rather than focussing on purely linguistic preoccupations. Instructions for example, exclude interventions through correcting or completing what candidates have to say[590]. Interactions between 'self' and 'other' should create two-way communication, broadly structured and facilitated by appropriate questioning and comment. As interlocutors, teachers should probe content, facilitating extended, intellectually-sophisticated and linguistically-expressive interactivity. In certain cases, failure to follow prescribed duties severely hinders authentically viable language-production. Candidates may be overly restrained through lack of occasion to perform adequately. This finding is reiterated in *Subject Reports*[591]. Teachers also bear ultimate responsibility in producing technically-satisfactory recordings, aiding unambiguous assessment and moderation by external, listening parties according to **IBO** procedure.

Studying anomalies in **IBO** criterion categorisations, differentiations and applications for assessment and evaluation illustrates the system's limitations. Thus, *Internal Assessment* topic choices must be amenable to appropriate presentation and discussion, facilitating performances capable of matching the most sophisticated criterion-descriptors. Certain topics and approaches make such attainment difficult, since presentations may alternatively, be too complex for adequate treatment at the chosen level, too banal to stimulate genuinely interested interchange, incorrectly set within an inappropriate genre (such as recitation of written discourse), or indeed seek closure in communication between candidate and teacher-assessor. These problems appear to affect few cases researched.

Similarly, time-prescriptions adversely constrain some performances, given topic choices. Candidates either require more for adequate development and discussion, or conversely too much is available for exploring overly-simple, superficial subjects. In anomalous cases, a few suggest that appropriate guidance on rubrics was lacking, or not fully heeded.

For certain candidates, microphone use and audio-tape recording appear to inhibit authentic production. Beyond concerns for technical quality, such activity requires psychological preparation. Students must be prepared to perform within the limitations of international contexts, where recourse to trained, external interlocutors is prohibitive for reasons of cost.

As further related, moderation procedures allow manipulation (or in extreme cases, disregard) of constraints on authentic language use. During the final year of a course, candidates are graded in *Internal Assessment* for tasks largely derived from classroom and individual performances, in contexts fully favouring continual, authentic assessment *per se,* as defined in Chapter 4[592]. Such productions are often influenced by familiarity gained through collaborative preparation with known teachers and within known school-based situations, thereby diminishing the effects of environmental constraints. Construct validity may be very high, although reliability limited by the inaccessibility of moderation as a means for re-assessing unrecorded texts.

Conclusions and *Written Production*

In *Paper 2*, the design facilitates authentic expression according to research criteria, albeit to a lesser degree than in *Internal Assessment*[593]. Restrictions of task-choice to one from prescribed lists of six forms a significant constraint, increased by choosing *Paper 1* readings as a base. This contrasts with the conclusions of Gipps, Linn, Linn and Dunbar and others, related in Chapter 4[594], yet stands at a polar opposite to less authentic, positivistic approaches whereby choice is entirely excluded as an immeasurable, subjective variable, increasing unreliability in evaluation and requiring for its elimination, valid comparability of all standardised task-constructs.

Initially, **IBO** candidates depend on sufficiently accurate reading competence for choosing suitable tasks and appropriately responding. However, rubrics and individual designs permit little scope for further negotiations of meaning, or continuous interaction through facilitation of extended, responsive, two-way communication, as necessary in fully authentic expression[595]. Nothwithstanding, *Chief Examiners* suggest that linking written tasks to readings from *Paper 1* has little adverse effect on response quality, whenever such tasks are chosen[596].

Further limitations on authenticity in **IBO** designs are clear. First and foremost, writing authentically for generally-defined readerships and in satisfying criteria for *Curricular* and *Pragmatic Authenticity*, demands appropriately written responses. These would not normally be transcriptions of oral responses for orally-based tasks, such as addressing an audience. Authentic task-specifications of genre are therefore essential. Writing for a readership also implies intention to 'publish' in some form (either openly or privately), whether readers are unknown and unresponsive or not. This requires demonstrable awareness that language-production is controlled by autonomous selves, prepared partially to surrender control to other readers who may use resultant texts for their own, possibly non-negotiated purposes. Alternatively, it implies recognition of capacities and occasions for 'repairing' misconceptions through written communication. Interactive interchanges are adapted and maintained with propositions from other and response by self, should readers be both known and engaged in relevant exchange. Meeting expectations for at least implicitly-designated readerships is a significant criterion for success, made explicit through many **IBO** task-designs, as well as through the relevant assessment scheme.

Traditional, pen-and-paper, point-in-time examinations, are evidently imperfect in propagating fully authentic expression. Certain unambiguous discrepancies and inadequacies in **IBO** task-design were noted, analysed and reported in Chapter 6. Indeed, the most significant illustrate disagreements for awarding and aggregating discrete marks in *Task* and *Message*, especially for cases where task-response is deemed irrelevant through inaccuracy in reading the requirements. (An instance was quoted for composing diary entries reflecting thoughts and emotions on leaving home in a near future, rather than recent past, from *Paper 2* in May 2001). From the examples, **IBO** criterion-categorisations may jeopardise authentic message-reception by examiners in their roles as interested readers.

Certain aspects of assessment-criterion design hinder regularity in applying procedure. Categorising *Task* and *Message* as a single criterion has been noted. Matters are obscured and even confused by separating this conflation from *Language* (and also from *Presentation*). Certain evaluations are difficult explicitly to justify under programme philosophy, aims and objectives. Through lack of alternative guidance and without the constraints of holistic evaluation, as inevitably required in assessing real-time, oral presentations and interactions, examiners may study written productions atomistically, word-by-word and structure-by-structure. They may resort to traditional, psychometric and linguistic comparisons with canonical norms of perceived, standard language, rather than assess authentically-communicated, expressive value, thus compromising the coherence of verdicts and their fit with programme philosophy. The predominance of linguistically-structural elements, evident in many of the *General Grade Descriptors*, indicate a significant influence, though these are only employed at the end of *Grade Award Meetings*, to check and balance through final, triangulating moderations of sampled responses, graded at each discrete level.

Whereas *Internal Assessment* designs may favour authentic expression by focusing evaluation on construct validity, though necessitating extended, interpretative moderation for acceptable reliability and credibility amongst 'stakeholders' in **IBO** programmes, the design for examinable *Written Production* allows easier control of reliability. This is at least so within each subset defined by all responses to any given task. Eliminating large numbers of teacher-assessors, as interactive interlocutors influencing the shape and course of candidate speech, and as raters in *Internal Assessments*, improves reliability at the expense of construct validity. In most cases however, the effects have not been found excessive, since reliability is predominantly established through recourse to standardisation and repeated, identical moderations of both candidate-productions and assessor-judgements. Criterion-referenced grade-boundaries are established anew in each examining session.

Sampled assessments are re-assessed by **IBO** examiner *Team Leaders* and *Chief Examiners*, the whole being subject to statistical manipulation, creating longitudinal records of individual assessor reliability, archived at **IBCA**. *Reports* clearly exclude any form of longitudinally-established, positivistic norm-referencing, as determinant in assessment and evaluation.

The Relation of Qualitative Assessment to Quantitative Evaluation

Throughout this thesis, constraints have been imposed on conceptualising authenticity through programme-design. In particular, problems arise in linking qualitative and relatively straightforward, if interpretative assessments of authentic language use, to quantitative evaluations. These are discretely grouped at differentiated levels of language mastery, eschewing non-interactive, atomised, 'objectivised' and positivistic methods of norm-referencing. This results in consensually-derived conversions of moderated descriptions of performance-based competence and attainment into numerical scores, forming grades in a seven-point range common to all **IBO** programmes, though for *Group 2 Languages,* communication levels are distinctly defined as *A2, B* or *Ab Initio*.

Hence the capacity of **IBO** designs to stimulate assessable, authentic expression through weighting criterion-referenced competences for final, quantified aggregations of measured language skill and knowledge, is important, if not centrally so. Totalising scores conflates variables requiring further investigation. The practice influences matching to programme philosophy, aims and objectives, its problems remaining unresolved. Detailed research scrutiny is required for greater purchase on the effects of distributing marks both discretely and in aggregations, as theoretically and practically realised by the **IBO**.

For *Language B,* distributions form a ratio of 70:30, divided respectively between external examination and partially-continuous, internal assessment. The system is rationalised as desirable for credibility and reliability in gaining broad acceptance of 'high-stakes' testing of this type. Such determinations of value partly indicate the influence, albeit indirect, of university and government representation within the **IBO**, heeded for enhancing international recognition of its certificates and diplomas.

However for equity if no other reason, establishing acceptable grade-reliability across different tasks for multiple administrations of assessment, is a fundamental concern. Threats to construct validity in task and rubric design need considering, since excessively artificial and unnecessary constraint may deform initiations of authentic expression, restrict language-production, and distort interpretations of quality in arriving at final judgements. Through compromise in attaining reliability, validity should be as little disturbed as possible. Ideally viewed, standardisation of papers, the choice of tasks for assessment, and evaluation of their products should be straightforward.

Idealised means for valid and reliable assessment require perfectly-consistent design, perfectly consistently evaluated in application through perfectly-consistent, interpretative moderation of criterion-referenced judgements of quality and effectiveness. In actuality, the inspection and possible editing of tasks through standardisation, interlocutor and assessor training, sampling and moderating grades, all are processes essential to ensuring appropriate administrative consistency, guaranteeing controls for equity in formal assessments offered to varied institutions and individuals in an academic 'market', as well as strategic co-ordination of the large numbers of people involved.

Assessment Criterion Categorisations

With such influences on authenticity in language use, investigations should not only consider equitable, justifiable weighting for discrete, skill-based components of an entire scheme, but also the point-value distribution of percentaged mark-allocations within each single-component criterion. Hence, *Internal Assessment*, valued at 30%, covers listening and speaking, communicatively evaluated under three distinctive, criterion perspectives: *Task-Response* and *Message, Interaction,* and *Language*, with the latter considered as competence and attainment in the command of grammar, vocabulary, idiom, and so forth. Despite a 10% maximum allocated to each criterion, the research evidence shows all criteria as mutually interacting, allowing implicit value transfers across discrete categories, permitting multiple assessment within a single criterion, albeit reprioritised through different criterion perspectives, yet creating interpretative difficulties for assessors.

Procedurally distinct from oral performances, *Written Productions* are unconstrained by simultaneous, real-time assessment. Criterion-categorisation may more explicitly refer to authentically interdependent features of language use, crossing criterion boundaries. Therefore, assessing *Language* implicitly, and occasionally explicitly, influences assessment of all three criteria, sometimes in combination of any two. With French, register is evaluated under *Presentation*, though given a defined readership, the category may refer to *Language* quality and appropriacy of *Task Response*. Clear examples are provided in candidate selections of simple or elaborated lexis, sophisticated usage of conditional tenses or subjunctive moods, consistent production of recognised, 'standard' forms of grammar, vocabulary and pronunciation and so forth. The theoretical, cultural and sociolinguistic foundations of **IBO** philosophy and programme aims appear in tension with practice. Assessors seem concerned in these cases with residual, psycholinguistic, purely structural and linguistic design-features, illustrated by positivistic tendencies in specifying criteria and procedure.

In addition, the problem of interpretative 'compensation' by assessors requires more thorough research, indicated as it has been by greater interrater-variance, when measured per criterion, rather than for aggregated totalisations of scores. Precisely-weighted, value justifications must be validated for all discrete criteria, and confirmed as reliable in minimally-distorting, qualitative description of authentic expression, once transformed into quantitative, numerical scores. Discrete evaluations must be capable of aggregation without compromise in original construct validity for any task proposed. They should individually accord with recommended philosophy, aims and objectives for authentic communication. For ethical consistency, teaching, learning, assessment and the ultimate evaluation of these should determine the style and content of a given system, complete with its associated assessment regime.

In this, the **IBO** programme studied may appear biased towards evaluations of written language, favoured by aggregations awarding 70% of final marks to displays of knowledge and skills in reading and writing. Only 30% are devoted to speaking and listening. Concomitantly, these skills are separated into major and minor components, devaluing the relative importance of listening. As revealed, the approximate percentage weighting of measurements of listening skill (as opposed to language-based skills of message presentation) may account for less than 10%. Hence, listening is only explicitly assessed under *Interaction* in *Internal Assessment*, though it is clear that any spoken message may be influenced, and even improved in overall quality through appropriate interaction with interlocutors. Such interaction evidently requires demonstrable listening skill, if on occasion, only implicitly so.

Furthermore, **IBO** *Group 2* categorisations as *A2, B,* and *Ab Initio Languages,* or as *Higher* and *Standard Levels* require supplementary research. The implications of such division, discretely conceptualising language standards, and differentially evaluating quality in authentic expression, need investigation. The research evidence has illustrated inconsistencies in programme design, most strikingly evident in the absence of definition of minimal performance at any level, commonly evaluated at a score of zero.

Further Unresolved Problems

For assessing and evaluating authentic language use, the perception of inauthentic constraint within the *French Language B* programme led to devising the key research questions. The quest for answers has revealed possible ways forward for achieving more satisfactory balances in validity and reliability, as compromises common to any system of assessment and evaluation.

At observed meetings, examiners suggested improvements of standardisation procedure. One recorded proposal required the inclusion of earlier examination scripts in each batch of papers sent for standardisation, with return to the **IBO** within a limited period of time, thus minimising possibilities for leaks of confidential examination material. The process would take place at least two years before publication of each examination paper. Such scrutiny could allow greater concentration on assessing appropriate task-differentiation for promoting authentic expression. Differential measurements of competence in mastering structural features of given language systems, discretely defined for given levels and groupings regardless of contextualisation by task, may thereby be avoided.

More significantly, the research leaves further issues unresolved, particularly those concerning socio-cultural contextualisations of assessed task-designs. Subsidiary problems concern both the relative 'difficulty' of different languages offered for formal assessment under a single scheme, and the production of candidates with different home-language substrates[597]. Evidently, these factors significantly influence authenticity in appropriately expressive communication.

Further complications and sources of invalidity are introduced through examiner familiarity with candidates' 'strongest' language. In particular, this allows 'easier' comprehension of 'message' through the sharing of specialised and 'unnatural' (in this sense 'inauthentic') 'interlanguages' between candidates and examiners. Such may favourably, if unintentionally, influence assessment under all criteria, compromising construct validity in task-based response for which particular audiences and readerships have been specified. Indeed for producing the most effective texts, the likely concerns of 'others' as listeners and readers, must be closely heeded.

Nevertheless, the broad concept of authenticity in language use seems more evidently problematic in *Paper 1* designs for evaluating comprehension, than for the more productive components of the **IBO** design. Indeed, the whole question of authenticity either as linguistic performance appearing 'natural' for 'native'-speakers, or as choice of reading material from the productions of the societies of such speakers, and adequately graduated by difficulty, (allowing range in assessment and evaluation), remains a source of ambiguity and difficulty. This is particularly evident from problems in the longitudinal standardisation of examination papers, across different examination sessions.

Possible Resolutions Indicated by the Research

Potential gains in procedural consistency, facilitating authentic communication for **IBO** design, assessment and evaluation purposes, have emerged. For example, *Paper 2* assessment criteria could be easier to apply should concepts of *Task* and *Message* be considered discretely. As in the experimental scheme, criteria for assessing *Task* as 'choice of genre in written production'; 'relevance'; 'convincingness' can be devised under unifying notions of authenticity. *Finder* and *User Authenticity,* and *Authenticity of Purpose* permit measurement of recognitions and the appropriate addressing of 'other'. *Message* may be recategorised within such conceptualisations, retaining notions of 'internal coherence', or acceptable 'flow', be it logical, imaginative, emotional, and so forth; relating specific, developed exemplification to relevant argumentation; indeed as 'having something worthwhile to say' and ' worthy of communication', in writing to a potential readership. Further Van Lier-derived categorisations of authenticity include such features. They allow the problems of linguistically-competent candidates who fail convincingly to communicate something relevant and worthwhile, more easily to be evaluated through criterion-referencing. Experimental investigation has highlighted paradoxes left unresolved by applications of **IBO** assessment criteria. Simultaneously, it has indicated possible resolutions, particularly under re-evaluation employing the Van Lier-derived criteria with 'plussages'.

In addition, graduations in criterion-description could be less ambiguous for assessor interpretation, especially for final evaluation *Grade Descriptors*. Tabulated commonalities and discrepancies were reported in Chapter 6. A clear example lay in the seemingly contradictory use of the French expression of "*maîtrise* limitée" (literally of "limited mastery") (*sic*) to describe a low level of achievement. Similarly, the categorical distinction of terms such as "moyen" (for "average") and "satisfaisant[598]" (for "satisfactory") could be improved, thus being easier for teachers, students, assessors and moderators consistently to interpret. Confusions in meanings apparently lead certain examiners, in traditional psycholinguistic and psychometric fashion, to emphasise purely linguistic factors, as provided by the *General Grade Descriptors*.

Possible resolution may lie in identifying three performance levels, as completed experimentally. These range from 'little', through 'adequate', to 'significant evidence supplied' per criterion, minimising likely contestation of differing assessor judgements, (though with two further categories added at the extremes, to identify either incontestable attainment or a total absence of evidence, as empirically-developed refinements). In the **IBO** model, ambiguity apparently facilitates determining final value by reference to linguistic competence as 'structural', rather than 'effective', in communicative, task-based responses.

The Criteria and Procedures for Awarding Grades

As reported, observations of *Grade Award Meetings* identified strengths of **IBO** moderation procedure ensuring concentration on construct validity as the stimulation of situated, communicative and authentic language use, besides bearing witness to the *Chief Examiners*' professional competence, care and thoroughness. Constant reference is made to the foundational value of the assessment criteria, published in the *Guide to the Programme*. This practice confirms the predominance of a grounded, *Language B* assessment philosophy, based in the avoidance of positivistic assessment methods, to favour evaluation by pre-established criteria, even though occasional recourse to longitudinal, statistical measurement was noted as introducing triangulating refinements in judgement, as if under norm-referencing[599]. The relegation of such techniques to the status of aids for 'stabilising' criterion-referenced evaluations was respected on all occasions researched.

Evidence of further strength was observed in *Chief Examiners'* interpretative approach, developed from deep, longitudinal experience as teachers and as examination-designers, examiners, moderators and final evaluators. This allows evaluators at *Grade Award Meetings*, credibly to enter possible scenarios suggesting explanations for the thought processes (both emotional and intellectual) of candidates, and thus to reinterpret, even if partly subjectively, the relative 'difficulties' of examination questions. The very possibility of such intersubjective communication is of course central to the ontology of authenticity, though ways for increasing reliability in assessments, have been previously indicated. Supplementary perspectives are provided, stabilising judgements derived from criterion-referenced evaluation of the significant features of authenticity in communication.

The pressures of meeting practical, examination deadlines, necessarily structuring the entire evaluation process and evidently risking variability in *Examiner* and *Moderator* performance through fatigue, were anticipated as significant constraints. However, the requirement to establish maximum possible assessment validity and significant reliability, is not unduly imperilled by such factors. To claim as much, does not diminish the recognised severity of loading on *Chief Examiners* at relevant times of the examining year[600]. With *French Language B* at least, one of the most popular programme choices for *Diploma* and *Certificate* candidates, the claim appears plausible, given the research evidence.

One evident weakness of criterion-referencing procedure lies in problems raised when insufficient numbers of candidate copies, coupled with doubts concerning the relative 'difficulty' of examinations, threaten validity through inconsistencies, especially in *Paper 1*[601]. However, restricting task-type formats increases predictability for any examination that is transparent in its design, assessment and evaluation criteria and procedures. Publishing detailed rubrics and explanations may facilitate routinisation of exam preparation in 'high-stakes' settings, negatively affecting validity for assessments of authentic expression in comprehension and written production. At *Grade Award Meetings* for *French Language B* of December 2000, and for *German Language B* of June 2001 however, it was observed that closely-scrutinised 'problem' cases, re-assessed by different examiners (totalling up to a maximum of eight consultations for a particular language-production), meant that ultimate verdicts remained credibly valid and reliable. At candidate request, recourse to appeal could further enhance such multi-perspectival credibility though further replications of procedure[602]. The increase in readership provided by multiple *Examiners* and *Moderators* may indeed be taken to respect criteria for full authenticity.

Further Prospective Developments for Future Research

From the investigations completed, the identification, collection, description and analysis of empirical evidence sourced both in moderation, evaluation and reporting procedures and in associated **IBO** documentation, has led to better grounding, and deeper understanding of authenticity as an operational concept, viable for use in evaluations of language use at any communicative level. The parameters developed and adopted as working definitions for identifying and assessing features of authentic language-production need no differentiation for adaptation to any given programme. In truly authentic schemes, differentiation is provided by the appropriate design and standardisation of tasks, with lower levels requiring relatively simple response, and higher levels increasingly sophisticated interaction. Further research development would permit more comprehensive understanding of emergent themes, promisingly instructive and viable as alternative means of assessment that they appear to be.

For example, minimum assessment criteria for differing levels within a single programme, that is, for scoring greater than zero at either *Standard* or *Higher Levels* should be devised, as recounted. Further research could improve understanding of **IBO** 'standards', implicitly theorised for each level, the differentiations within *Group 2 Languages* programmes being analysed as a whole, clearly to distinguish scoring requirements throughout the range, for *Standard* and *Higher Levels* in

A2 and *B* programmes, and for the *Standard Level* only of *Ab Initio*. The *Group* may then be viewed under organisational philosophy, as a unified, interlinking range with five, categorical subdivisions[603].

Furthermore, for improving the valid generalisability of findings, the understandings, attitudes and practice of examination-designers, examiners, moderators and evaluators, should be collected and analysed in representative samples. In particular, data should relate to the problems of applying authenticity as conceptualised, in assessable categorisations to formal productions. Under research experiment, supplementary triangulation of results is desirable for more-securely founding the propositions of such data-manipulation.

The Experimental Research

Following Van Lier's work, the conceptualisation and categorisation of authentic language use has produced insights into advantages and disadvantages in applying such criteria to formal assessments and evaluations of linguistic production. For example, through integrating the evaluation of *Task, Message, Presentation* or *Interaction* and *Language,* aberrances threatening construct validity have been highlighted and made less ambiguous.

However, the limited reliability of research experiments, from recourse to a single assessor for devising and completing assessments, requires improvement through replication by teams of others and across a range of languages. The verdicts of different assessors, following Van Lier's evidently overlapping criterion-categories, yet based in different cultures and with different first languages, needs deeper understanding. Viability and in particular, the effects on point-values of weighting component aggregations for final evaluation need further investigation.

In practice for example, the distinction between *User Authenticity* and *Authenticity of Context,* as defined, is difficult to describe. Similar problems bedevil understanding of *Authenticity of Interaction, Intrinsic, Existential* and *Creator Authenticity* when addressing the concerns of self, *Authenticity of Purpose, Autotelic Authenticity*, and so forth. Further research of issues in weighting and aggregation is evidently desirable. The dilemmas resemble **IBO** attempts reliably and quantitatively to measure authentic language use through attributing described, weighted values as categorised, to language-productions in any language, whilst retaining the qualitative validity offered by the **IBO** model for assessment and evaluation.

Weighting criterion-values for validating aggregated assessment of individual performances by discrete criterion, is thus emphasised. As in the **IBO** model, experiment fails to separate discrete, qualitative descriptions from totalised, quantitative evaluations, for deriving overall meaning as numerical report on valid and reliable accumulations of quality. Further development and refinement of the research instrument is required for more practical exploration of key research issues.

The effects of comparing candidate topic-choices for oral assessment could thereby be more deeply considered. Assessment under all categories experimentally specified is crucially dependent upon such choice. The categories of *User Authenticity* and *Authenticity of Context* for example, are differentiated by mode of communication, be it oral or written. The implications require more detailed investigation.

Notwithstanding, experimental assessments have illuminated the advantages of a model designed for determining validity through triangulation, for understanding the coherence and consistency of **IBO** programme philosophy, aims and objectives, and for use in conjunction with assessments derived from the **IBO** scheme. In the context of authentic production, the experimental criteria refocus assessment attention on the significance of subject matter chosen for presentation by 'self', and on the ensuing interaction with 'other', as listener or reader, and subsequently as interlocutor or replier. The model is thus advantageous in more clearly identifying problems in performing for authentic communication. Purely structural concerns of positivistic, linguistic and psychometric measurement are eliminated from the design. The Van Lier model thus 'frees' assessors from attending to uniquely language-based qualities of candidate productions, stressing instead in culturally-situated interactions, the existential concerns of phenomenology together with those of sociolinguistics.

The model is also 'free' from specificities in language variety, however categorised. That is, assessing productive use by learners of any language, whatever unique, structural complexity there may be, is not distinguished *per se*. It is also 'free' of categorisation as 'foreign' (*Language B*) or 'second' (*Language A2*). Differentiation is measured by sophistication and language range in designing and standardising assessment tasks that require appropriate productive response, addressed to specified audiences or readerships[604].

The advantages of Van Lier's model include addressing problems of construct validity, posed for example by cases of reading aloud, as in sampled, *Internal Assessment* presentations. Likewise in *Written Production,* the model with its rubrics and descriptors for *Finder*

Authenticity and *Authenticity of Context,* can identify and assess plagiarism of content and language, be it from copying examination material proposed, or from memorisation of the work of others, in searching for 'model answers'. Exclusively positivistic measurements of accurate reproduction of predefined lexis, grammatical and idiomatic structures, become irrelevant to the assessment criteria and procedures, whether behavioural, or non-interactive and intersubjective methods are retained for teaching and learning, or not.

Indeed, Van Lier highlights the need for 'purpose' and 'motivation' (for 'autotelia') in 'successful' linguistic interchange. Language reception and production are intimately integrated with personality and personal identity. The model simultaneously allows different types and genres of expression to be assessed in common, being applicable to both oral and written modes, to different text genres, and to the use of any standard language, creole, dialect, patois, or even personally-based 'interlanguage', through its focusing on the assessment of exchange in interactive dialectics of communication between 'self' and 'other. It emphasises dynamic, personal development in social and linguistic interactions, thus underscoring the 'existential' aspects of authenticity, as defined by philosophers such as Sartre. Fundamentally, it focuses attention on assessing situated abilities to express self to others through the medium of language, and to continue that expression in intersubjective, dialectical interchange.

CHAPTER EIGHT

THE PREMISES OF THE RESEARCH

The Aims

Theory and practice have been linked through researching non-positivist, criterion-referenced, interpretative assessment and evaluation systems promoting authenticity in modern, foreign-language teaching and learning. Ideally, descriptions and graduated measurements of productive language interactions should refer to explicit statements of programme philosophy, aims and objectives. For **IBO** *Diploma Languages B*, qualitative and quantitative assessments, matching performance descriptors to numerical evaluations, are meaningfully and mostly consistently correlated by construct validity. Experimental research suggests means for reducing inconsistencies and anomalies in designing and applying appropriate criteria. Given research methods and evidence, greater critical insight into the structure, purpose, processes and products of the *French B Standard Level* programme has been achieved.

Traditional understandings of authentic language use have been found ambiguous. Explicit, conventional definitions derived from specialist literature are difficult to apply in benchmarking for assessment. Transparently justifiable, valid, reliable and credible evaluations, relating as fully as possible to the authenticity of task-based responses, need enhancement. Searching for possible improvements has been a major aim.

When identifying and categorising components of authentic communication through discretely-graduated criterion descriptions, refined, theory-based definitions emphasise the complexity of performance in situated, linguistic relationships. In fulfilling general research aims, it was possible to indicate and explore potentially reliable measures of valid, task-based assessments, stimulating authenticity in 'target' foreign language use.

Throughout, authenticity has been closely linked with relationships established and expressed both reflexively within 'self', and communicatively between 'self' and 'other', in pedagogically and culturally-determinate contextualisations. For selecting materials and specifying pre-determined social roles, purposes and functional settings in task-design for formal assessments, 'being authentic' is not solely a matter of criterion categorisation and description, representing a world beyond the self of individual examination candidates with varying degrees of realism. Following Van Lier, authenticity integrates states of individual awareness and autonomy, situated with a given socio-cultural and linguistic *milieu*. It is a necessary product of purposeful and meaningful interaction between individuals, sharing something in common.

In this view, **IBO** assessments and evaluations highlight the complexity and fluidity of the linguistic interchanges they promote. Quality judgements are based in part, not on positivistic comparisons with authoritative, pre-defined norms, but on the subjective impressions and personal interpretations of listeners and readers, simultaneously working as professional assessors and moderators. Grasping the nature of subjectivism in assessment and evaluation is central to understanding the processes of authentic language use. It allows detailed consideration of many contestable elements of communicative interactivity between listener and speaker, reader and writer. Through focusing on questions of validity and reliability, theoretical criteria for authenticity and their application in practice receive critical scrutiny. In formal settings, many of the contestations described are thus resolved.

Candidate performances, alternatively assessed under **IBO** rubrics, criteria and procedures and experimentally within a paradigmatically-different model applied with similar procedure, have produced data for triangulating **IBO**-based assessments. The central problem of identifying assessable features of quality in authentic expression, has been emphasised. Greater relief in understanding has emerged, aiding evaluation and appreciation of selected components of the *Group 2 Languages* programme, and thus serving a major goal.

Simultaneously through exploring a known scheme, bridges between theory and practice have been established. The **IBO** programme selected is well documented. Its products have been researched with longitudinal sampling of data from varied, recorded, task-based performances of a large number of candidates over a range of assessment administrations, including that of their assessors and moderators. With authenticity as a vantage, congruencies, similarities and differences emerge on comparing **IBO** 'theory in practice' with an 'espoused theory' of authentic language use, rendered 'practical' through experiment.

Notwithstanding the complexity of definitions and processes investigated, greater consistency and precision in determining the components of authentic expression have been attained. Furthermore, the qualities identified have allowed viable specification of criteria for valid assessment, even though the relation of qualitative judgements to quantitative evaluations in statistically significant ways remains problematic and limited in reliability, given recourse to a sole, identical researcher and rater.

Most evidently, the experimental criteria require refined simplification, for more precise application to significant samples of language-production, across a range of languages and assessed by a range of examiners and moderators, both specifically trained for the purpose and otherwise. The assumptions underlying the nature, relevance and assessability of authenticity in language use may be thus further explored, made explicit and tested for validity and reliability in formal evaluations. This research aim has therefore, only partly been fulfilled.

The Objectives

The project design has linked general aims with specific major objectives. In summary, the research emphasised questions of validity in devising standardised tasks that require authentic response within set rubrics, and permit qualitative, criterion-referenced assessment, consistently correlated with quantitative evaluation. Procedural reliability in determining the overall qualities of examination performances was also of central concern. Investigating use of authenticity as conceptualised by the **IBO** for its examination designs, standardisations, assessments, moderations and evaluations, served these objectives and applied to oral and written productions of French in representing a significant *Diploma* grouping and level.

The objective of improving theoretical benchmarking, was served by investigating understandings of 'authenticity' through analysis of publications and selected documentation produced for internal, administrative purposes of the **IBO**.

The objective of integrating the philosophy and general aims of the *Group 2 Languages Programme* as a view of pedagogy and learning, with practical assessments of its products, allowed tentative generalisation of 'grounded' understandings of authenticity in its various guises, developed from analysis of the work of selected examination-designers, standardisers, internal and external assessment candidates, examiners, examination-moderators and evaluators. Inconsistencies in comprehension and practice were identified through comparing the varying, emergent definitions, understandings and usages, both explicit and implicit.

Similarly, evidence for collection, analysis and discussion, contextualising theory and practice and reported in Chapter 6, founded responses to the following questions:

- In what ways and with what effects did the assessments studied 'position' the following:
    - The **IBO** as an institution?
    - selected candidates of assessment and evaluation?
    - the moderator and examiner whose assessments and reports were analysed?
- What problems were identified in candidate language-productions, attributable to institutional and procedural inconsistencies sourced at the **IBO**?
- What implications did identified problems and inconsistencies have for applying assessment procedures?

Discussion of the design leads to conclusions as to whether the research objectives have been validly and plausibly attained.

The Design

In effect, the research-design refocuses assessment interest, turning away from traditionally pre-eminent reliance on positivistic, psychometric, or purely linguistic approaches to validity and reliability in measuring language-performance, towards more social, intersubjective, interactive and hence authentic schemes. This alteration emphasises listener and reader perspectives and judgements in the assessment process, integrated within the roles of assessor and moderator. In **IBO** sampling and processing *Internal Assessment* and *Written Production* data, the individual views of *Teacher-Assessors, Internal Assessment Moderators* and *Assistant Examiners* are highly significant. They should ensure adequate construct validity through respecting the criteria of authenticity in language use, with acceptable reliability established through re-assessment and moderation, including recourse to further listeners and readers, as *Internal Assessment Moderators* and first-line *Examiners,* as well as to the judgements of *Team Leaders*, *Deputy* and *Chief Examiners*.

Experiment has shown that differentiated, graduated, qualitative assessments of authentic language use are validly convertible to meaningful, quantitative scorings. In a large number of cases, the outcomes closely correlate the **IBO** scheme with an experimental model, devised for assessing and evaluating the authenticity of the language-productions researched. Reliability measurements suggest acceptable matching to **IBO** criteria and procedure, whilst closely respecting requirements for assessing authentic language use, as categorised and developed from theory.

The results obtained under either system appear at least equally informative. With experimental work, greater emphasis on authenticity facilitates valid and reliable evaluation of 'aberrant' examples of task-response. For interpreting and applying **IBO** assessment criteria, such examples gave rise to various, significant difficulties, leading to contestable results. In this context, the experimental model reduces the incidence of 'problem' cases, through more appropriate criterion-referencing and measurement. It is evident however, that not all problems of assessment and evaluation may thus be eliminated.

In interpreting research claims, there is nonetheless, a need for caution, since limitations in research design and application are clear.

One significant constraint reducing the effectiveness of comprehensive, multi-dimensioned study of the *Group 2 Languages* programme, is the limitation of detailed analysis of task-design and assessment-criterion categorisation to data for *French Language B*. Given this restriction, scope is further restricted through emphasising authentic production at *Standard*, rather than at *Standard* and *Higher Levels.*

Thus scrutiny of significant boundaries and interfaces, not only between differing levels within a particular scheme, but also between the three discrete, *Group 2 Languages* programmes, was cursory. In assessment and evaluation, these evidently overlap in range (albeit implicitly), and in linguistic 'level'[605]. For a single *Language A2, B* or *Ab Initio* under a common philosophy, comparing the range of discrete programme aims and objectives, assessment task-design, standardisation, criteria, application, moderation and final evaluation, with sampling and analysis of relevant productions, was therefore little investigated. Likewise, little data from a common programme and level were analysed for comparing performances across a range of differing languages.

Hence, restrictions in available resources for completing comprehensive investigations removed possibilities for enhancing the claim that the research-design allows precise delineation of boundaries, cohesive inspection of areas overlapping discretely-delineated subject levels, and comparison of evidence sourced in alternative, though similar domains. The production of validly and reliably-analysed data, from which more broadly generalisable, though still meaningful and useful conclusions could be drawn, was more narrowly and more invariably circumscribed.

In particular, the research implies a need for appropriate, and confident justification of weighting decisions applying in criterion-referenced assessment, in order fully to respect construct validity when aggregating discrete, qualitative assessments as holistic, quantitative, final evaluations. This remains only partly addressed. More comprehensive consideration of weighting issues is desirable for further research into the validity of aggregated evaluations, and their outcomes after moderation and the determination of point-score boundaries in grade attributions, as numerical representations of performance quality. Indeed, the design has highlighted cases appearing either typical or anomalous in assessments of authentic language use, both with respect to **IBO** criteria and procedures, and to the model employed for producing experimental triangulations. Full conclusions thus remain tentative. Whilst it is claimed that anomalies may be reduced in incidence through greater attention to issues of authenticity, they cannot completely be eliminated, removing evidence of problems of categorisation in describing authentic assessment criteria.

Further issues of validity and reliability raised by the research design stem from inevitable constraints imposed by limitations in material resourcing for completing comprehensive investigations. The reliance on an individual researcher and practitioner involved in all the processes investigated is evidently restrictive, as well as a point of strength. On the one hand, recourse to a single assessor and moderator may favour greater consistency. For producing assessment data, this is especially so in the case of longitudinal measurements completed over a three-year time span. Stable understandings derived over time from a single individual (albeit necessitating good faith in operating under such assumptions), enhancing the likelihood of minimally-variable replication of procedure and avoiding necessary standardisation through training, appear adequately reliable. For the researcher, **IBCA** reliability checks over the same period of time indicate such assumptions as reasonable, even if fine-grained detail has inevitably been lost in the sampling and statistical processing involved. Use of a single researcher, assessor and moderator eliminates needs to control further variables, introduced by multiplying the number of such personnel, desirable though this may be.

On the other hand, significant sampling of understandings and applications of the experimental assessment criteria and procedures, across a representative range of raters and rating activities is also thereby excluded. Besides procuring the services of such raters, their training in the design and processes of experimental evaluations, and employment in the relevant context, additional controls of assessment validity and reliability would inevitably be required. Such lay beyond the possibilities of the research.

Detailed investigation of authentic language use in relation to different, teaching and learning approaches influencing assessment and evaluation was also excluded. This is regrettable, given the completion of some data-collection for researching self-determined choice amongst student approaches to assessment-task selection, response preparation, composition and final editing. In any context, such features have been deemed essential for promoting authentic language use. In the literature, consideration of the intentional effects of 'washback' on pedagogy and learning, or of particular understandings of key components determining value within any assessment and grading system, are generally recognised as significant. They influence teacher and learner motivation to conform and excel within the prescriptions of a given system, restoring purposefulness in linguistic performances as a fundamental component of being authentic.

Reliance on willing, open and comprehensive provision of relevant, unpublished documentation, archived at **IBCA**, is an item of trust, ideally requiring control in any comprehensive design for research. Evidently, such documentation is intended for confidential, internal use in examination contexts where security must constrain and limit possibilities for the most blatant effects of 'washback': be they as plagiarism, or unfair advantage in the preparation of formally-assessed responses[606]. In the eventuality, and self-evidently, the **IBO** fulfils a role as gatekeeper for accessing much of the documentation utilised, though the research-design has required no breach of confidentiality, with anonymity respected in all significant cases. Indeed, as an academic institution working in close liaison with universities and supporting research of this type, the **IBO** fulfils by present example, its claims to favour the search for, and production of this kind of new knowledge[607].

Lastly, negotiated access for further investigation of the processes by which consensus is established within examination-designer groups, supplementing the evidence provided by the documents researched, may be seen as desirable. Such lay beyond the project's scope.

Procedure and Practice

The research illustrates weaknesses and strengths in integrating theory and practice within a multi-dimensional model. Method included an initial formation of hypotheses and rationales founded on prior experience, both personal and professional, of a single researcher. To this was added a grounded approach, creating interfaces between theory and practice. Besides identifying, collecting, describing, and analysing the discourse of relevant **IBO** documentation, an eclectic mixing of qualitative and quantitative strategies extended empirical investigation. These may be recalled as:

- grounded analysis and comparative evaluation of sampled, *French Language B, Standard Level, Internal Assessment* oral productions, and written productions for *Paper 2*;
- observation and recording of *Grade Award Meetings* for moderating and evaluating examination scripts;
- discourse analysis of sampled, **IBCA** documents, both formal and informal, intended for internal, **IBO** use in devising and administering examinations.

Furthermore, the whole was placed within a loosely-evolving framework, originally inspired by the approaches of *Action Research*. That is, discrete exercises of earlier research gave rise to progressive refinements in overall design and form for instruments employed.

Further sources of significant evidence were identified, facilitating improved data-analysis, the results being partly influenced by increasingly comprehensive consultation of relevant, theoretical literature. The research proceeded in cycles whose major punctuations were simultaneous with deadlines for *Internal Assessment Moderation* and *Assistant Examining* in *Diploma Programme* examination sessions, held annually in April and May. The completion of data-collection and analysis at the end of each 'cycle' led to amendments and further development of model and method for the ensuing stage.

Research circumstance ensured that this procedure, if evidently practical for a part-time researcher, remains intrinsically problematic. Greater longitudinal control of data-collection and production for the assessments completed would have improved the generalisability of conclusions, given reliance on the understandings and interpretations of the single researcher and the ever-present possibility of variance in judgement over time. Stability and consistency appear ultimately as assumptions: as invariables, they are perhaps reasonable, if not indeed fully reassuring. Substantiated by both limited and informal cross-checking during data-analysis, and external, **IBO** evaluation of annual, reliability measurements of employee performance as professional assessors for the organisation, an acceptable degree of consistency may validly be assumed. Indeed, searching for absolute reliability, independent of ever-changing experiences anchored in the contingent flow of time of all such processes, resembles a search for a Holy Grail, given the qualitative, and ultimately interpretative and subjective basis of all assessor judgements, under which criterion-referenced and authentic assessment must take place. In this respect however, it may be noted that no data indicating significantly contrary concerns, emerged from the cycles of research.

Whilst it seems feasible to draw general conclusions from experimental findings and apply these to assessment practice in formal contexts for 'high-stakes' purposes *per se*, it should be recalled that such was not a research aim. Nor, for practical reasons, could extensions be included. These require supplementary data-collection, notably from comprehensive sets of alternative assessments for further triangulation of validity and reliability claims, made with identical instruments, applied under verifiably-similar procedure, though completed by representative teams of trained and 'standardised' raters.

Indeed, as a post-script in evaluating experimental research procedure, further investigation of authentic language use in assessment requires scrutinising choice as an explicit examination rubric for respondents.

The research instrument devised requires further refinement in design to increase construct validity and permit further assessment of variables in oral and written productions, dependent upon:

- virtually unlimited variety in candidate-determined choices of subject in oral presentations for *Internal Assessment* under the existing **IBO** scheme;
- the effects on authentic language use of such freedom of choice;
- candidate determination of task selected from a possible and prescribed choice of six, for examinable written productions;
- authentic language use in responses to prescribed topics from *Paper 2*, relating to themes from pre-read texts, invariably provided as required reading for *Text-Handling.*

Such choice, determined by candidates themselves, clearly questions equity in comparability for assessment under commonly-specified criteria and procedures in examination-design and standardisation. Whilst the experimental instrument devised may obviate needs discretely to control for the effects of such variables, through detailed specifications of categories of evidence under *Curricular, Pragmatic, and Personal Authentication*, candidate, teacher and assessor attitudes and practices should be surveyed for further validation through alternative triangulation of the research data. Such survey did indeed form a component of the original proposal, but was excluded for shifting the research focus towards issues of teaching and learning in preparation for assessment, albeit in association with the potentially significant effects of 'washback'.

The Viability of the Research

In evaluation of the design, final conclusions concern the plausibility and generalisability of research findings.

The research has investigated claims to validity for a given programme, examining whether **IBO** practice establishes what is claimed in theory. In this, findings are generally positive, though **IBO** understandings of the key concept of authentic language use remain ambiguous. In assessment and evaluation, such ambiguity creates evident anomalies, albeit as exceptions to a general trend. Experimental manipulation deepens insight into their nature and suggests ways for reducing their incidence, whilst retaining existing task-design and assessment procedure for evaluating performance.

Simultaneously, reliability in evaluating *French B* language-productions was scrutinised. Assessments can be replicated by different raters across time to give acceptably similar results, despite the anomalies described. However, high reliability is in itself dependent upon *Assessor, Examiner* and *Moderator* training for conformity in interpreting assessment criteria and applying uniform procedure. It is also dependent upon reliable sampling under a timetable with exacting deadlines. Repeated assessment and the re-moderation of such assessment are required as reliability checks at each sampling point. Possible error and inadequacy are recognised as irradicable, though increasingly unlikely as procedure unfolds, with candidate appeals permitting further replication of the entire process in a final check.

The **IBO** system relies on the central process of moderation by trained personnel, based in sampling assessments by *Internal Assessment Moderators* and *Assistant Examiners*. The noted problems of regularity in interpretation may thus recur, with plausibility in outcome resting on the prior knowledge, experience and integrity of *Team Leaders, Deputy* and *Chief Examiners*, co-ordinated by the **IBCA** *Subject Area Manager* and *Director of Assessment.*

In this, observation of moderation and grade-awarding by *Teacher-Observers* ensures transparency and integrity, with comment recorded in unpublished official reports. Through statistically-significant samplings of candidate, assessor and moderator work, finely-tuned, evaluation adjustments are possible, given recourse to limited, normative, longitudinal comparison across two (though not more) administrations of identical assessment sessions. The apparently inconsistent contrast with the **IBO**'s published commitment to criterion-referencing in evaluation is partially compensated by replications of assessments over any single session, with any single language-production receiving the individual attention of five or more assessors and moderators. It is further compensated by fine gradation in criterion grading across each component and across aggregated results, first by applying task-specific criteria to each formal assessment exercise, and then by frequent reference to the *General Grade Criteria,* during moderation. Further checks occur under regular review and system updating, subsequent to *Examiner* reporting. In organisational terms, the **IBO** requires full programme review at five-year intervals, for each of its programmes.

Construct validity and reliability for criterion-referenced, and fundamentally interpretative systems are also enhanced through the regular training of teachers, as *Assessors* and *Moderators*, whether as **IBO** employees or otherwise. Furthermore, the employment over time of relatively stable teams of *Assistant Examiners, Internal Assessment Moderators, Team Leaders, Deputy* and *Chief Examiners* allows statistically-valid and reliable replication of assessment and sampling procedures, and historical data-collection by **IBCA**, concerning individual rater-reliability.

Thus, despite a research-design reliant on the work of a single rater and researcher, threats to reliability may partially be compensated by the researcher's professional experience as an **IBO** employee, by regular retraining and exposure of individual understandings to critical appraisal from teachers, both familiar with and new to the system at **IBO** training workshops, and ultimately by replications of assessment and sampling procedures for controlling rater interpretations over time, throughout the life-span of the programme researched.

From relevant description, analysis and discussion, outcomes appear plausible. Triangulation with **IBO** results, generated latitudinally by a range of assessors, and longitudinally across examining history, confirms impressions. They are also generalisable, though within the limited bounds of the research parameters, where certain key areas remain ‘fuzzy’ and occasionally ‘problematic’, requiring contestable interpretation. Given validity for **IBO** assessment criteria with reliability for assessment procedure, interpretative cautions being noted, the research results illuminate questions of situated authentic language use.

As a project in itself, the research results appear plausible, for coherent and consistent triangulation with **IBO**-produced data is demonstrable, albeit with limitations. Given validity for the experimental criteria and their use, the restricted, though confirmed reliability of assessments, seems generalisable, albeit with reiteration of the provisos already noted. Comprehensive features of authentic language use may be discretely categorised as criteria for assessment and evaluation. The results permit critical description, analysis and discussion of the philosophy, aims and objectives of the **IBO** in its *Diploma Programme* for *Group 2 Languages.*

NOTES

PART I

CHAPTER ONE: Early Hypotheses

Initial Approaches and Rationales

[1] This pilot project formed the content of the *Stage 1*, preliminary proposal and research of the **Open University**'s programme for the *Doctorate of Education*, presented to the university in April 2000, and approved as a design for fuller research from September 2000.
See Israel, (2000), *op. cit.*

[2] From this point onwards, this organisation will be labelled by its commonly-used acronym, as the **IBO**.

[3] These terms are derived from Argyris and Schon (1974), *op. cit.*

[4] The present study however, is limited mainly to French, as a major exemplar of a discrete domain for **IBO** *Group 2 Languages*. Further, progressive focussing of the research narrows investigation to the *Diploma Programme* for *French Language B*, with comparative references to other domains and levels defined within this group, both for the *Internal Assessment* component, and for examinations as *Paper 2*, or *Written Production*, at *Standard Level*, in particular.
See Chapter 2.

[5] All approaches attempt to denote, explain, assess and evaluate aspects of language use, whether in reception or in production.

[6] As will be seen in Chapter 2, these are defined as second, or 'foreign' languages, ranging in level from beginner up to full bilingual equality with a first, 'mother', or 'native' language.

[7] In the context of this project, 'assessment' will be taken to refer to the essentially qualitative process of derivations of value through the formation of judgements that match with descriptive criteria, whereas 'evaluation' will be taken to refer to the transformation of such qualitatively-based judgements into quantitative representations and numerical scorings.

[8] 'Positioning' is a concept derived from the work of Fairclough (1989), amongst others. In the context of the research, being 'positioned' is taken as a constraining effect of both the form of the curriculum, assessment and evaluation systems of the relevant **IBO** programmes, and the pedagogical approaches, selections and practice of teachers associated with preparing candidates for **IBO** assessment. It also includes considerations of validity in meaningful, published evaluation for prestigious, 'high stakes' awards that frequently determine access to further education, (or quite simply and more generally, for culturally-approved, social prestige).

The Origins of the Hypotheses

[9] These interests were further developed and systematised through study at the **Open University**, following courses and completing assignments for the degree of *Master of Education*, from 1995 to 2000. Pilot projects on the theme of authenticity in second language pedagogy and learning were developed within this framework.

[10] The career biography of the researcher is relevant to the development of the research and its further rationale. He is a full-time teacher of the programme concerned at the *Istanbul International Community School,* Turkey – a non-denominational, co-educational, not-for-profit, independent international school, governed by its parent-body, and serving English-speaking students from the international community, aged 3 to 19, and following the three major programmes of the **IBO**: the *Primary Years' Programme,* the *Middle Years' Programme* and the *Diploma Programme*. Other than English, the language of most instruction, French is the only language to be taught within these curricula.

[11] Fuller details are given in the concluding section of the subsequent chapter. See p. 67

[12] See *Note No. 1.*

[13] In particular, these took place during observation, assessment, moderation and evaluation exercises of the **IBO**'s *Grade Award Meetings* for *French* and *German, Languages B,* in December 200 and June 2001, respectively.

A Preliminary Understanding of Authenticity

[14] The term is taken from Csikszentmihalyi (1990), who contrasts it with 'autotelic' purpose, defining it as activity that seeks reward other than within its own enactment. See Csikszentmihalyi (1990), *op. cit.*

[15] This section of the **IBO** will henceforth be labelled by its commonly-used acronym, as **IBCA**.

[16] These include for example, candidates with identified, special learning difficulties such as those experienced by the blind.

[17] In this context, any grounded conceptualisations of authenticity are subsumed within the general parameters of a communicative philosophy of language acquisition and of the assessment and evaluation of such acquisition, explicitly adopted by the **IBO**.

[18] Such inconsistency and incoherence may be taken as intimately related to bounds established by the **IBO's** *internally designated* constraints, and apparent from the vantage point of **IBO** use of 'authenticity' as a key concept for guiding functional aspects of communication in a foreign language. This is acknowledged to be both embedded in the linguistic culture of all users of language, and the rationale for the measurement and validation of attainment under the relevant assessment scheme.

The Definition of Key Questions

[19] The research questions initially developed from the pilot project, were framed as follows:

- What understanding of authenticity emerges from analysis of the **IBO**'s publications, and to what use is this concept put by the organisation?
- What grounded understanding of authenticity emerges from analysis of the work of selected examination candidates?
- What grounded understanding of authenticity emerges from analysis of the work of a selected examiner?
- What inconsistencies in understanding and practice can be identified through comparison of the varying definitions and usages, both explicit and implicit, as outlined above?

These questions were refined to account for improvements in understanding central issues as the research progressed, and to profit from access to new, or previously unplanned, yet relevant sources of data. The guideline was to establish greater precision of focus, depth of field and fitness for purpose, rather than alter perspectives or change direction in the progress of investigation.

Refinement and Development

[20] Devising such an exercise in triangulation deepened understandings of authentic language use, as a concept in itself. Simultaneous, grounded analysis of evidence established the significance of this conceptualisation, in situation within relevant **IBO** programmes and examinations.

CHAPTER TWO: The Organisational Context

Preface

[21] This information is based on major **IBO** publications as referenced, and in particular on the general brochure *The **IBO**: Education for Life*, **IBO** (2001e), *op. cit.*

It also uses material available to the public on the organisation's website at: www.ibo.org

Further details on the background of the organisation are given in **Appendix 1**.

The International Baccalaureate Organisation

[22] The origins of the **IBO** date from the establishment of the **League of Nations** in Geneva, Switzerland after the First World War.

[23] See **IBO** (2001e), *op. cit.*

The **IBO**'s constitutional and legal status are summarised in **Appendix 1**.

[24] See for example: **IBO** (2001e), *op. cit.*

[25] As an example of the scope of the organisation, these numbered 3,700 in May 2001.

[26] See **IBO** (2001g), *op. cit.*, p. 6.

[27] They communicate in face-to-face meetings, by telephone, fax, letter, email and dedicated internet discussion fora in order to achieve common understandings of duties and co-ordinate activities.

[28] As a further example of the scope of arrangements, the team is currently composed of 31 *Chief Examiners*. To this number may be added a *Chief Assessor* for the non-examined *Theory of Knowledge* component of the *Diploma Programme*.

[29] See **IBO** (2001g), *op. cit.*, p. 6.

The IBO *Diploma Programme*

[30] This information is based on major **IBO** publications as referenced, and in particular on the general brochures: *Guide to the Diploma Programme,* **IBO** (1997a); and *The **IBO**: Education for Life*, **IBO** (2001e), *op. cit.*

No notable differences in content have been discovered in consulting such documentation, separated by an interval of four years in publication, unless otherwise stated in discussion.

The research also uses material available to the public on the organisation's website at: www.ibo.org

[31] In certain documents, this adjective is replaced by "demanding".
See **IBO** (1997a, 2001e), *op. cit.*

[32] See **IBO** (2001g), *op. cit.*

[33] See **IBO** (1997a, 2001e), *op. cit.*

[34] See **IBO** (2001e), *op. cit.*

[35] See **IBO** (2001e), *op. cit.*

[36] The other relevant domain groupings may be noted and summarised as follows:

- *Group 3* or *Individuals and Societies*, consisting of social sciences such as business and management, economics, geography, history, Islamic history, information technology in a global society, philosophy, psychology, social and cultural anthropology;
- *Group 4* or the *Experimental Sciences*, consisting of biology, chemistry, physics, environmental systems, design technology, with practical laboratory work and a complementary emphasis on "moral and ethical

issues and a sense of social responsibility [.....] fostered by examining local and global issues";

- *Group 5* or *Mathematics and Computer Science*;
- *Group 6* or *The Arts*, consisting of visual arts, music and theatre arts "with emphasis placed on practical production [.....] and exploration of a range of creative work in a global context".

In addition, various *School-based Syllabuses* may be authorised as alternatives for given 'subjects' in *Groups 2, 3, 4* and *6*, especially for the purpose of meeting any relevant and particular, national requirements for students of this age.

Alternatively, a *Group 6* 'subject' may be replaced by a second choice from *Groups 1* to *5*.

As stated, the programme also requires the satisfactory engagement with three interdisciplinary elements, in the form of the following:

- *Theory of Knowledge,* with at least 100 hours of teaching time "intended to stimulate critical reflection on knowledge and experience gained inside and outside the classroom, [.....] question[ing] the bases of knowledge [for an] aware[ness] of subjective and ideological biases, [for developing] the ability to analyse evidence, [for] appreciat[ing] other cultural perspectives, [for] reflect[ing] on all aspects of [students'] work throughout the programme, [and for] examin[ing] the grounds for the moral, political and aesthetic judgements that individuals must make in their daily lives" and leading to the production of written essays and oral presentations;
- *Creativity, Action* and *Service*, with a recorded expenditure of time in which the "whole person" is educated "to help students become responsible, compassionate citizens", with emphases on "shar[ing] [.....] energy and special talents with others [by] develop[ing] greater awareness of themselves and concern for others, and the ability to work co-operatively with other people". (Examples of such engagement are given as theatrical or musical production and community service activity);
- *Extended Essay*, with the production in approximately 40 hours of private study, of a 4,000 word research paper from a very wide range of more than 60 subject options, in either the *Group 1* or *Group 2 Language*, as chosen by the student, and that requires the investigation of a topic of special interest, intended to "acquaint diploma students with the kind of independent research and writing skills expected by universities".

In this, the organisation demands balance in curricular provision, to encourage internationalism through the integration of language skills and knowledge in more than one language, employed and practised both in and outside the classroom, and requiring experience and reflection across the entire range of components chosen in any authorised programme of study.

See **IBO** (2001e), *op. cit.*

The articles of the programme require authorised **IBO** schools to register candidates for either a full *Diploma*, or a selection of individual subject *Certificates*.

It is expected that these schools schedule formal instruction for a minimum of 150 hours in a *Standard Level* subject, and a minimum of 240 hours at *Higher Level*, over the course of the two years devoted to the curriculum.

In many cases, this time allocation and prescription forms the major (and sometimes only) criterion that clearly distinguishes requirements for *Standard* and

Higher Levels. For the present research, the recommendation and its implications are further discussed in succeeding sections of the present chapter.

Schools partaking in **IBO** programmes must be formally authorised by the organisation and evaluated in five-year cycles under a range of criteria. When and where necessary, this authorisation is withdrawn. In general, such an eventuality occurs only for reasons of unsatisfactory administrative practice by the school concerned. No judgement of student 'quality' is ever made in determining the removal of authorisation from any particular school.

See **IBO** (1997a), *op. cit. Article 16,* pp. 25 -26.

[37] See **IBO** (2001e), *op. cit.*

[38] The choice should range through all the subject groupings, to the exclusion of none.

In addition as had been noted, satisfactory completion and submission of work in *Theory of Knowledge, Creativity, Action and Service,* together with an *Extended Essay*, is also obligatory.

See *Note No. 36.*

There are also various conditions for acceptable combinations. These do not concern the research, though it may be noted that they discourage duplication in study across the six 'subject' domains. The intention is to ensure that candidates participate in the full range of offerings within the *Diploma* design, according to the philosophy, aims and objectives of the organisation.

Such conditions evidently do not apply in the case of students presenting work in individual subject domains for the purpose of discrete certification.

See **IBO** (1997a), *op. cit.,* pp. 19 –26.

[39] Various 'failing conditions', relating to the minimum point score required in each combination of components are also outlined in the articles of the *General Regulations*. In themselves, these do not concern the present research and hence are not reiterated here.

See **IBO** (1997a), *op. cit. Article 9,* p. 22.

[40] Point-in-time, external examinations formally take place in May or November of the second year of instruction in the programme. The large majority of candidates for the May examination sessions are located in the northern hemisphere, and those for the November sessions, in the southern hemisphere. However, in each case, there will be some who are retaking examinations in order to improve scores from a previous session.

[41] It will be seen later that the *Internal Assessment* for *Group 2 Languages* forms one exception to this general rule, since the form and content of the assessment rubric remain constant across all examining sessions at a given level.

[42] It is organisational policy to require a ratio of between 20% and 50% as internally-assessed work, with no less than 50% as production under supervised, examination conditions. In this way, the overall design requires a guarantee of incontestable authenticity for the bulk of the material presented as a candidate's own work.

Information communicated verbally to the researcher, by the **IBCA** *Director of Assessment* in August 2002.

The Principles of Moderated Assessment and Evaluation

[43] See **IBO** (1997a), *op. cit.,* p. 14.

The following general description outlines **IBO** understandings and policy in this respect:

> "each student's performance is measured against well-defined levels of achievement consistent from one examination session to the next. Grades reflect attainment of knowledge and skills relative to set standards that are applied equally to all schools. Top grades are not, for example, awarded to a certain percentage of students."
> See **IBO** (2001e), *op. cit.*

[44] This procedure is more fully reported in **Appendix 2**

[45] See **IBO** (1997a), *op. cit.,* p. 14.

[46] In the case of languages, this refers to tests of "fluency, command of vocabulary, grammar and structure in a taped exchange with a language examiner [.....], called the oral component of the examination"
See **IBO** (1997a), *op. cit.,* p. 14.

[47] See **IBO** (1997a), *op. cit.,* p. 14.

[48] In the case of *Group 2 Languages, Language B, French* and *German*, this moderation and evaluation process is reported in detail, subsequent to observation, in **Appendix 2**.

[49] These are made available to the **IBO**'s authorised schools both in hard copy and via the Internet.

[50] It should be noted that the *General Grade Descriptors* are only made available to a wider public by these means.

[51] This grading system is common to all Groups and subject areas of the *Diploma Programme*.

[52] The relationship between these various tables, at first glance apparently reworking similar data, is not made explicit in the documentation published. Their significance is as a control for distortion in the aggregation of component scores and grades, and as such is discussed in the conclusions of Chapter 7.

It may additionally be noted from semi-structured interview with the **IBCA** *Director of Assessment* in August 2002, that *General Grade Descriptors* serve as a final point of reference in grade-awarding, ensuring that procedure has led to broadly consistent verdicts. They are not used to determine assessments per component, subsequently to be aggregated into the final total score and grade.

[53] See **IBO** (1997a), *op. cit.,* p. 16.

[54] Information supplied to the researcher by the **IBCA** *Director of Assessment*, August 2002.

Given the limitation of research to components relating to language production, detailed investigation of final grade award criteria and procedures has not been undertaken.

[55] See **IBO** (1997a), *op. cit. Article 9,* p. 22.

In 2001, the **IBO** could in this way claim that over 40,000 students had been internationally assessed, with a success rate of approximately 80% for the award of the *Diploma*, an apparent constant since 1997 when nearly 30,000 students were assessed.

See **IBO** (1997a). *op. cit.*, p. 17, and **IBO** (2001e), *op. cit.*

Language Groupings in the *Diploma Programme*

[56] The availability of a particular language at examination depends largely upon the demand communicated to the **IBO** by schools with potential candidates.

(Informal communication to the present researcher by the **IBCA** *Director of Assessment,* in August 2002.)

[57] However in one example, 'Netherlandish' (termed 'Dutch', and composed of various Dutch and Flemish dialects grouped together in a single, homogenous 'official' language, defined in content and usage by the appropriate national and linguistic authorities) may be noted as differentiated from Afrikaans, for which separate provision is made.

See **IBO** (1997b), *op. cit.*

[58] This is documented in relation to a large number of languages in the *Language Specific Annexe to the Language B Guide.*

See **IBO** (1997b), *op. cit.*

[59] **IBO** (1997b), *op. cit.*, p. 10.

[60] **IBO** (1997b), *op. cit.*, p. 10.

[61] Examples of varying attitudes are frequent, ranging from the "relaxed attitude towards the spoken form" of Afrikaans that permits regional variation "provided that the context is appropriate"; to the acceptance of "deviation from standard pronunciation, standard negation rules or rules for case endings" and "lexical variations from different dialects" in the case of Arabic; to the production of examination papers in "traditional and simplified characters" in the cases of Cantonese and Mandarin; to the encouragement to respect new, governmentally determined revisions in particular of orthography in the cases of Dutch and German; to respect of the "regulations of the **Academy of the Hebrew Language** in the case of Hebrew; to more detailed and interesting statements of the situations pertaining to Bahasa (Indonesia) and Norwegian, where it is stated for the former that:

> "It is essential to open students' minds to [.....] differences [between 'dialect' and 'official' language] to avoid the very real danger that students will only be able to communicate in a one-way direction."

In the case of the latter, respect of governmental determinations is made explicit, as follows:

> "Bokmål is usually the variety of Norwegian taught as a foreign language and has therefore been chosen as the main language for *Norwegian B*. Nevertheless, the *Language B* programme is based on authentic material and should reflect the diversity of the language.
> For this reason *Paper 1: Text-Handling* will include mostly texts in Bokmål, but one text in each text-handling paper will be in Nynorsk. This reflects the proportion of Nynorsk compulsory on Norwegian television."
> **IBO** (1997b), *op. cit.*, pp. 8 – 9.

Moreover, certain non-national languages such as Welsh and the classical languages of Greek and Latin, conventionally recognised as 'culturally-homogenous' through forming a discrete, assumedly rarely-contested standard, are also included by the **IBO** in the range available within the *Diploma Programme*.

[62] **IBO** (1996b), *op. cit.*, p. 3.

The term 'native' is further defined in a footnote as follows:

" 'Native' in this context refers to the language acquired by a speaker through exposure to it from an early age. It is normally, or has normally been for an extended period, the language of the speaker's home environment. Related terms are 'mother', 'first', 'home'."
See **IBO** (1996b), *op. cit.*, p. 3

However, it may be noted that in the setting of many international schools, the majority of which in practice, use instruction through the medium of English, this becomes a student's *A1* language. In fact it may well be a 'second', rather than a 'native', 'mother', or 'home' language.

[63] See **IBO** (2001e), *op. cit.*

[64] **IBO** (1996b), *op. cit.*, pp. 3 – 4.

[65] **IBO** (1996b), *op. cit.*, p. 4.

[66] This feature of programming is described in greater detail, analysed and discussed in succeeding chapters.

[67] **IBO** (1996b), *op. cit.*, p. 4.

The *Language B* Programme

[68] It should be noted at the outset, that the published, **IBO** documentation on which this section is based refers largely to English-language versions. Although for present reporting, the relevant French-language versions have been taken into

consideration, they do not to diverge to any significant extent from the original, English versions on which they are based.

IBO (1996b), *op. cit., passim.*

[69] This is by a wide margin. Entries at *Standard Level* predominate, with English, French and Spanish by far the most numerous language choices, and with entries in most recent years totalling over four thousand in each case.

For the May 2002 session of the *Languages B* programme, 20,648 candidates entered examination, of whom 1,247 were at *French, Higher Level,* and 5,142 were at *French, Standard Level.* Spanish attracted an approximately 20% larger entry, with other large entries represented by English and German.

[70] **IBO** (1996b), *op. cit.*, p. 4.

[71] **IBO** (1996b), *op. cit.*, p. 4.

[72] **IBO** (1996b), *op. cit.*, p. 4.

It is further noted that:

> "students with limited learning experience of the target language or those with no previous learning experience of the target language, but who live in a country where the language is spoken, may be able to follow the *Language B* course at subsidiary level."
> **IBO** (1996b), *op. cit.*, p. 4.

[73] The relevant sections are entitled: *Nature of the Subject: Language B; Aims; Objectives;* and *Syllabus Outlines.*

See **IBO** (1996b), *op. cit.*, pp. 6 – 23.

[74] The relevant aspects of these statements are presented, analysed and discussed at appropriate junctures in later chapters.

[75] **IBO** (1996b), *op. cit.*, p. 6.

[76] **IBO** (1996b), *op. cit.*, pp. 9 – 10.

[77] **IBO** (1996b), *op. cit.*, p. 6.

[78] Such understandings have been made explicit in the latest editions of the *Guide to the Programme*: *Language B,* where "authentic materials" are defined as "spoken or written, printed or electronic materials that have been produced to satisfy the needs and expectations of native-speakers of the target language".

IBO (2002b), *op. cit.*, p. 13.

[79] Indeed, at *A2* level, this feature is now explicit in a statement under the subheading of *Classroom Environment*, that:

> "Teaching must be provided in the target language, and learning should be placed in the contexts that prepare the students for actual use of the language."
> **IBO** (2002a), *op. cit.*, p. 15.

For *Language B,* the latest requirements state that *inter alia*:

> "teachers should aim to provide a typical monolingual environment where teaching is provided in the target language and learning is placed in a context that would be familiar to speakers of that language."
> **IBO** (2002b), *op. cit.,* p. 13.

In this respect, it should be held in mind that the latest statements of the organisation represent an evolution of the programme, rather than a change of approach, and that the range of language levels offered form a single continuum, or "spectrum".

See **IBO** (2002a, 2002b), *op. cit.*

In this context, statements relevant to the *A2* programme, whilst not the focus of present research, represent perhaps the most unambiguous declarations of the **IBO** in its conceptualisation and use of 'authenticity', colouring the more ambiguous use of the notion at 'lower' levels, such as those for *Language B.* Indeed the border between the programmes is intentionally ambiguous, with teachers exhorted to place students "appropriately" to represent an "adequate challenge" for learning, and avoiding the "amass[ing of] points in an educationally sterile fashion".

See **IBO** (2002a, 2002b), *op. cit.*

[80] See **IBO** (1996b), *op. cit.,* p. 6.

The 'weighting' of discrete items of language skill will be specified, analysed in its effects, and discussed in a subsequent section.

[81] See **IBO** (1996b), *op. cit.,* pp. 6 - 7.

[82] See **IBO** (1996b), *op. cit.,* p. 7.

[83] **IBO** (1996b), *op. cit.,* p. 8.

[84] The detailed objectives for measurement are listed and reported in the *Guide to the Language B Programme* at both *Higher* and *Standard Levels.*

IBO (1996b), *op. cit.,* pp. 9 – 10.

[85] **IBO** (1996b), *op. cit.,* pp. 9 – 10.

[86] See **IBO** (1996b), *op. cit.,* p. 10.

[87] These are further defined as those typical of communication through the medium of a target language in the context of university study, such as attending lectures, participating in seminars, tutorials and practical work, independent reading of literary and non-literary works, writing notes, essays and reports.

See **IBO** (1996b), *op. cit.,* p. 9.

[88] See **IBO** (1996b), *op. cit.,* p. 9.

[89] **IBO** (1996b), *op. cit.,* p. 10.

[90] **IBO** (1996b), *op. cit.,* pp. 11 - 23.

[91] **IBO** (1996b), *op. cit.,* p. 11.

Exemplars are given both of detailed themes and of text types, though it is emphasised that first, the study of themes should not be an "end in itself", and that in practice, the categorisations are so general as to include most materials available in some form or other. See **IBO** (1996b), *op. cit.,* p. 12.

[92] **IBO** (1996b), *op. cit.,* p. 11.

[93] See **IBO** (1996b), *op. cit.,* pp. 13 – 23.

The Structure of Assessment in *Language B*

[94] At this juncture, arrangements are outlined, with more detailed discussion following in the succeeding chapters devoted to the presentation and analysis of empirical evidence. The data provided here is summarised from relevant **IBO** documentation.

See **IBO** (1996b), *op. cit.,* pp. 24 –25.

[95] **IBO** (1996b), *op. cit.,* p. 26.
See also **IBCA** (2001c), *op. cit.*, p. 5.

[96] In this, 'authenticity' refers to texts sourced from publications in target languages, intended for 'native'-speaker readerships and unadapted in task-settings.

[97] As discussed in subsequent chapters, the provision of well-defined, but restricted options from which a candidate may select a single task for response, should be noted for its impact on issues of authenticity in language use. Suffice it to state at this juncture, in anticipation of the presentation and discussion of the following section, that concern is shown in the *Instructions* provided for examination designers, to ensure a practical minimum of constraint in examination settings.

[98] In this context, the relevant *Subject Guide* states:

> " 'Oral work' should be understood to comprise both the productive skill of speaking and the receptive skill of listening".

One of the aims is:

> "to allow listening skills to be integrated into the oral component".
> **IBO** (1996b), *op. cit.,* p. 27.

[99] See **IBO** (1996b), *op. cit.,* p. 27.

[100] Exemplars of appropriate activities are provided in the programme *Guides.*
See **IBO** (1996b), *op. cit.,* pp. 32 – 33.

[101] Examples are provided in **Appendix 4.**

[102] Some effects of this design are described and analysed, with discussion in Chapter 6.

Examination Design and Standardisation

[103] This documentation identifies protocols for the design and standardisation of the relevant examinations and is both temporarily confidential and normally, solely available for the internal use of the **IBO**. However for the present research, a selection was made available to the researcher. Access to **IBCA** archives was granted in July 2001.

With no evidence to the contrary, the documentation consulted was taken to represent a comprehensive sample. Moreover, the necessity for the strictest confidentiality with material referring to the 2000 and 2001 examining sessions was evidently diminished, once the administration of the examination, its assessment, moderation and evaluation were complete, with results in the public domain, and once the defined time-limits for queries and appeals had lapsed.

[104] The archival documentation was annotated, with data selected as relevant to the description, analysis and evaluation of criteria, procedures and practice in this area. The results provide an example of evidence relating to the design and standardisation of the examination session in *French Language B* for May 2001, though in the absence of conflicting evidence from other sessions, this has been taken as typical. The notes taken from this source were shown to the *Director of Assessment* on completion, and photocopied for the *Subject Area Manager* for reasons of ethical consistency, respect for the confidentiality of the material they contain and acknowledgement of the intrusive nature of this aspect of the research. They were not subsequently edited or in any way altered by **IBCA** personnel, and are hence deemed accurate as a representation of content and procedure.

[105] For May 2001, the *Centralised Examination Paper Production* section of **IBCA** (**CEPP**) produced 850 examination papers, together with corresponding mark schemes, via communication with more than 250 external examiners.

[106] *Op. cit.*, p. 4.

[107] The role of the latter is defined explicitly as: "essential [.....] in ensuring the academic integrity of **IB** assessment within each subject. [*Subject Area Managers*] are involved throughout the process of examination paper production, providing guidance to examiners and other members of the team to ensure that the question papers take account of the nature of the **IB** candidature and are a fair and appropriate reflection of the **IB** programmes which they aim to assess." (*Op. cit.*, p. 4).

Imposed constraints on the design of the examination papers by official *External Advisors* are identified in paper-specific instruction booklets and concern the following:

- "layout, internal rubrics, spacing of questions, line numbering, etc.
- For all language examinations the general instructions to candidates are provided by **CEPP** [the *Centralised Examination Paper Preparation* department of **IBCA**] in English, French and Spanish on the front cover of the examination paper according to a standard format." (*Op. cit.*, p. 6.)

[108] In the case of less commonly-examined languages, the *Standardiser* is provided with translations of the proposed examination from the target language into one of the working languages of the **IBO**: English, French or Spanish. This practice, through introducing an element of potential distortion, be it linguistic or cultural, has clear implications for issues of authenticity that are partly considered in Chapter 6.

However, with the delimitation of the research restricted to *French, Language B*, such a factor is of no direct concern. All communication relating to the production of examination papers in this domain are drawn up in French, as a working language of the organisation.

[109] See **IBCA** (2001c), *op. cit.*, p. 5.

This is understood in the context as material produced for native-speaker readers and audiences for purposes that concern neither explicit processes of acquiring language for its own sake, nor the assessment of such language acquisition.

A discussion of the effects of such editing for *Paper 1* and the related implications for *Paper 2* in the single case of the optional task based on the reading material of *Paper 1*, is to be found in Chapter 6.

See pp. 163, *et sq.*

[110] This detail concerns factors relating to legibility on reproduction in an examination booklet, the respect of legal issues concerning copyright, and of ethical issues concerning marketing and the use of specific items with a function as commercial publicity. As such, these relate to the production of *Paper 1: Text-Handling* for the May 2001 examination and hence do not concern the central discussions of the research.

[111] This follows in later chapters.

[112] See the **IBO**'s *General Instructions* to examiners responsible for the design of specific papers.

See **IBCA** (2001b), *op. cit.*, p. 9.

[113] These are for *Higher* and *Standard Levels, Paper 1: Text-Handling* and *Paper 2: Written Production.*

See **IBCA** (2001c), *op. cit.*

[114] Only the instructions relating to *Paper 2 Written Production* are of direct relevance to the concerns of the project, and duly reported here. The form and content of *Paper 1 Text-Handling* are relevant insofar as the general themes presented may be related to the tasks posed in *Paper 2*. Nevertheless, the instructions define these relations as no more than "tenuous", and hence, this aspect may be discounted at this juncture.

In the case of *Paper 1*, questions of authenticity may be taken to relate to the manipulation of the linguistic material presented to candidates in order to ensure conformity with the discrete rubrics set by the **IBO** for this particular examination paper.

Discussion of this follows in Chapter 6.

See pp. 163, *et sq.*

[115] **IBCA** (2001c), *op. cit.*, p. 4.

[116] **IBCA** (2001c), *op. cit.*, p. 4.

[117] **IBCA** (2001c), *op. cit.*, p. 4.

[118] **IBCA** (2001c), *op. cit.*, p. 4.

[119] **IBCA** (2001c), *op. cit.*, p. 5.

[120] **IBCA** (2001c), *op. cit.*, p. 5.

[121] **IBCA** (2001c), *op. cit.*, p. 1.

[122] **IBCA** (2001c), *op. cit.*, p. 1.

[123] **IBCA** (2001c), *op. cit.*, p. 1.

[124] **IBCA** (2001c), *op. cit.*, p. 4.

[125] **IBCA** (2001c), *op. cit.*, p. 4.

[126] **IBCA** (2001c), *op. cit.*, p. 5.

[127] **IBCA** (2001c), *op. cit.*, p. 5.

[128] **IBCA** (2001c), *op. cit.*, p. 5.

[129] Further details and discussion follow on this point, in the subsequent section.

[130] Original emphasis indicated in bold type.
IBCA (2001c), *op. cit.*, p. 5.

[131] **IBCA** (2001c), *op. cit.*, p. 5.

[132] Original emphasis indicated in bold type.
IBCA (2001c), *op. cit.*, p. 6.

[133] Further analysis and discussion of this aspect of the examinations is provided in Chapter 6.

[134] It should be held in mind that the latter are explicitly privileged as the primary source from which the form and content of the examinations and their assessment criteria are derived.

[135] The nature of this 'sophistication' remains implicit and unexplained.

[136] They may be understood as relating to the difference defined in the *Subject Guides* that examination at *Standard Level* is appropriate after 150 hours of study in a teaching programme for the relevant component of the *Diploma Programme*. The equivalent figure given as appropriate for *Higher Level* is 240 hours.

See **IBCA** (2001c), *op. cit.*, p. 4.

[137] **IBCA** (2001c), *op. cit.*, p. 5.

Investigation of claims such as these is presented and discussed in relation to the data selected and described in Chapter 6.

[138] **IBCA** (2001b), *op. cit.*, p. 1.

[139] **IBCA** (2001b), *op. cit.*, p. 1.

[140] **IBCA** (2001b), *op. cit.*, p. 1.

[141] See *Note No. 114.*

[142] Researcher's italicisation and emphasis.

[143] A putative example is given, as noted in the previous section. The implications are discussed in relation to the empirical evidence of the research, presented in Chapter 6.

[144] **IBCA** (2001c), *op. cit.*, p. 3.

The ambiguity in the definition of distinctions between *Higher* and *Standard Levels* require further inspection and discussion with regard to their implications for issues of authenticity, as follows in later chapters. They need not unduly concern the presentation, preliminary analysis and discussion of documentary data concerning the **IBO**'s internal protocols. Indeed, it may be noted that the evidence for such ambiguity has been found primarily located within the selection, description and manipulation of candidate work, produced, assessed and evaluated solely at *Standard Level.*

[145] For ethical reasons, respect of the security concerns and interests of the **IBO**, and in conformity with agreements under which primary data from the organisation could be selected and collected, the notes from which this section is derived, were shown on completion, to the **IBCA** *Examination Papers Office*, or **CEPP** department, and to the *Director of Assessment*, with an invitation to comment. No alterations were made and permission was granted to make use of their content in the present report. A photocopy of the full set of notes was produced and passed to the **IBCA** *Subject Area Manager* for *Group 2 Languages.*

Assessment and Examination Administration

[146] **IBO** (1996b), *op. cit.*

[147] **IBO** (1996b), *op. cit.,* p. 44.

[148] The influence of structuralist linguists such as Halliday and Hasan as the source of 'authority' and subsequent 'justification' for the categorisations adopted by the **IBO** may be supposed as implicit, and as identifiable for those familiar with the work of the aforementioned analysts, presented and discussed in Chapter 4.

See Halliday (1975), *op. cit.,* and Hasan (1989), *op. cit.*

[149] **IBO** (1996b), *op. cit.,* p. 44.

[150] **IBO** (1996b), *op. cit.,* p. 44.

[151] **IBO** (1996b), *op. cit.,* p. 44.

[152] See *Guide to Language B*, **IBO** (1996b), *op. cit.*

[153] **IBO** (1996b), *op. cit.*, p. 37.

[154] Again, the influence in devising such a categorisation of written language may be related to the work of Halliday and Hasan, as noted previously.

See Halliday (1975), *op. cit.*, and Hasan (1989), *op. cit.*

[155] **IBO** (1996b), *op. cit.*, p. 37.

[156] **IBO** (1996b), *op. cit.*, p. 37.

[157] **IBO** (1996b), *op. cit.*, p. 37.

[158] The recommendation has been noted from other *Internal Moderators* and *Examiners*, and has been advocated in **IBO** training workshops organised for training teachers as assessors within the framework of the programme.

[159] **IBO** (1996b), *op. cit.*, p. 35.

[160] **IBO** (1996b), *op. cit.*, p. 35.

[161] See **IBO** (2001a), *op. cit.*

[162] These include administrative irregularities, the suspicion of fraud, failure to respect the rubrics of the examination or the detail of the tasks set, and so forth, and as such, do not concern the main body of the research as typical cases.

However, where comprehension of rubrics and tasks is at stake for appropriate response, serving as a stimulus to the authentic use of language, such is identified, analysed and discussed in Chapter 6.

See **IBO** (2001a,b,c), *op. cit.*, pp. 1 – 2

[163] **IBCA** (2001a), *op. cit.*

[164] **IBCA** (2001a), *op. cit.*, pp. 15 and 16.

The Weighting Values of Listening and Speaking, Reading and Writing

[165] The ratio of 70% : 30% for external examination and internal assessment thus falls within the overall requirements of the **IBO** for its *Diploma Programme*, though no explicit rationale for the distribution is published. In general, attempts are made to balance in appropriate compromise, the theoretically-strong reliability of point-in-time examination with the theoretically-strong validity of 'continuous' internal assessment.

Information supplied in conversation with the researcher, by the **IBCA** *Director of Assessment* in August 2002.

[166] See for example, the explanation of the role of *Chief Examiners*, given by the *Director General* of the **IBO**, conjointly with the *Chair of the Examining Board*, the latter representing "all *Chief Examiners* and representative *Chief Examiners*", amongst others:

> "*Chief Examiners* appointed from the universities and colleges bring with them vital outside influences, important to curriculum development and assessment procedures for the *IB Diploma Programme*. They also provide vital links with the institutions that take most of our students, post-diploma. This is of major benefit, in particular university recognition of the diploma."
> See **IBO**, (2001g), *op. cit.,* p. 6.

[167] See for example, **IBO** (2001e), where it is stated that these figures possess "international authority in their fields."

[168] See **IBO** (1996a), *op. cit.*

[169] This is stated in the relevant *Examiners' Manual.*
See **IBO** (2001a), *op. cit.,* pp. 1 – 2.

It may further be noted that cohesive linkage, a desirable criterion of *Presentation*, is explicitly defined in linguistic terms as "grammatical and lexical". Implicitly, the description and evaluation of qualities categorised under *Criterion C: Language* are repeated in this way, thus emphasising their importance.
See **IBO** (1996a), *op. cit.,* pp. 40 – 41, and 43 – 44.

[170] See **IBO** (1996a), *op. cit.*

[171] Examples of linguistic features considered under *Presentation* are explicit for "register", but implicit in concerns for language sophistication (understood as the establishment of appropriate register through the use of conditional forms for reasons of politeness, for example).
See **IBO** (1996a), *op. cit.*

Internal Assessors, Examiners, Moderators and Evaluators

[172] Indeed, a rationale for the requirement is given in the *General Instructions for the Moderation of the Internal Assessment Component.*
See **IBO** (2001d), *op. cit.*

It should be further noted that only the French version has been consulted.

[173] See **IBO** (2001d), *op. cit.*

[174] See *Examiners' Manual, Part 4,* **IBO** (2001a), *op. cit.,* pp. 4 – 5.

[175] See **IBO** (1997a), *op. cit.,* p. 15.

[176] See for example, **IBO** (2001a), *op. cit.*, *Section B, paragraph 3.2.*

[177] See for example, **IBO** (2001a), *op. cit.,* pp. 4 – 5.

[178] See for example, the joint statement of the *Director General* and *Chair of the Examining Board*:

> "Chief examiners are recruited mainly from universities and colleges throughout the world [.....] The **IBO** has in its examining board an international group of high expertise, covering the whole range of the curriculum, contributing to the well-being of its *Diploma Programme*. The distribution of *Chief Examiners* is not as uniformly spread through the regions as it might be, but a variety of measures are being taken to identify suitable people from all countries and to successfully recruit them as and when positions become available."
> **IBO** (2001g), *op. cit.*, p. 6.

PART II

CHAPTER THREE: Significant Theory: a Literature Review

Preface: Authenticity as Theory and in Practice

[179] This includes the devising of authentic tasks, designing and applying assessment criteria, moderating the results and attributing value by grades to the language elicited.

Preparing a Review of the Literature

[180] From analysis of empirical data, theoretical features of authentic language use are experimentally identified and evaluated as exemplars of authentic communication in practice.

[181] Albeit with 'second' and 'foreign', modern languages, and to a degree that may prove significant, a primary linguistic and cultural socialisation is assumed already to have taken place during acculturation and acquisition of a 'first' language, 'native', or 'mother' tongue. Further discussion of this point follows in subsequent chapters.

[182] In particular, this philosophical domain is defined in two of Sartre's major works, to which reference is continually made: *Being and Nothingness* (1946a), and *Existentialism and Humanism* (1946b), though these have been consulted in their original, French language versions.

See Sartre (1946a; 1946b), *op. cit., passim.*

[183] The significance of this in educational research is stressed for example by Scott and Usher, in whose work a plea is made for recognition and due consideration of the ontological underpinnings of all epistemology. The choice of ontological foundation is inherently loaded with value, thus highlighting the ethical and axiological concerns that must invariably accompany any such choice.

See Scott and Usher, (1996), *op cit.*

[184] Consequently, (though perhaps surprisingly for a philosopher initially employed as a school teacher), issues relating to methods and content in teaching, learning and assessment, and subsequently to the relations these imply between teacher and student, characterising institutions and their members, as well as 'societies' at large, are not examined. However, all are presented as integral constituents for the creation of authentic and viable, socio-cultural relationships.

[185] Subsequently, experimentally-developed criteria have been employed as structural items in the design of a research instrument for categorising, describing, analysing, rendering significant and discussing empirical data, thus illustrating language use in this context. In application, they afford means for consistent measurement, creating alternative perspectives for interpreting data and permitting multi-dimensional triangulation as a method for validating data produced under other assessment systems. In this exercise, samples of language assessed and evaluated under existing, **IBO** criteria are compared with assessments based on identical data, developed in the course of research.

Equally, they may give rise to evaluations closely linking 'espoused theory' with 'theory in practice', and entailing the final, 'justifiable' attribution of numerical scores to the outcomes of descriptive, criterion-referenced assessments. The strengths and deficiencies of the existing **IBO** scheme, as well as of experimental research, may thus become more apparent through such a multi-dimensional approach, stimulating sound, critical appreciation of the whole. Within its bounds however, it is assumed that the simplification inherent in the necessary process of compressing raw data to produce conceptual frames of reference neither grossly distorts, nor misrepresents original evidence. Nor, more importantly at a theoretical level, should it pervert the overall line of reasoning, sophistication or fine detail in Sartrean thought.

Traditions in the Philosophy of Authenticity

[186] See Heidegger (1927,1962), *op. cit.*

[187] Such represents a very brief reformulation of Sartre's preliminary conception of existential phenomenology as an assertion of the foundational perception that "existence precedes essence".

Put crudely, this states the awareness that "I am alive, and here", upon which knowledge all other awareness, or knowledge must be built. Without such awareness, there can be no meaningful formulation of definitions of existence.

See Sartre (1946a and b), *op. cit.*

[188] Any search for an *a priori* cause or purpose that is capable of representation, is held in Sartrean phenomenology to be 'inauthentic', in that such representation would constitute the product of an *a posteriori* search for rationalistic 'objectivity' of knowledge and understanding, allowing representation to be created in the first place. Any such search, for Sartre, is conducted by subjective entities who subsequently impose rational conceptualisations for 'explaining' the state and nature of the consciousness of individual existence on the existent, in order to define it. In other words, *I* can only 'know' and 'understand' after the fact of existing and *in temporal continuation* of that existence.

See Sartre (1946a and b), *op. cit.*

[189] One goal of materialism is the definition of aims containing and describing *a priori,* the phenomena of consciousness and of its expressions in thought as fixed, and fixable through the ultimate investigations of detached, empirical observation. Such phenomena may be analysed through intellectual processes, rigorously applied in schemes of scientific rationality. In this sense, the currents of empiricism, rationalism and positivism, evident in behavioural psychology, psycholinguistics and the associated psychometrics of individual and socio-cultural representation and verification, flow counter to existential and phenomenological affirmations that the bounds of thought, as well as of its expression, are limitless. Thought and its expression are both continually coming into being, and simultaneously evolving within the socio-culturally and temporally situated context of all being.

[190] It is similar to Heidegger's, summarised for example by Mills (1997), with references to Guignon (1984, 1993):

> "For Heidegger, authenticity is a uniquely temporal structure and a process of unfolding possibility. It is a state of being that is active, congruent, contemplative, dynamic, and teleological - an agency burgeoning with quiescent potentiality (Guignon). As such, authenticity is the process of becoming one's possibilities; and by nature it is idiosyncratic and uniquely subjective."
> See Mills, (1997), *op. cit.*

[191] It is not significant that reflexive, mental operations for organising and transforming selected perceptions into the internal representations of memory may be absent (as in purely sentient awareness), thus rendering irrelevant to 'self' any ordering for creating meaning in the apprehension of externalities. Indeed, infinite freedom in choice may appear limited through 'self'-chosen acts, trying to reify the objects of perception. These are attempts to stabilise *objects* of focussed attention as unchanging, despite their location in time. (For Sartre, the process is delusional). The same is not true when attention is internal, focussing on the phenomenon of 'self', within itself, since such thought is also located in time, and is influenced by the passage of time through the effects of memory. Hence, authentic situations of conscious perception are unstable and ever-changing.

[192] The socio-cultural and psychological consequences may of course vary in importance for 'self'. Indeed, from this line of reasoning, it may be understood that should rejection of assessment and evaluation results prove common practice, the socio-cultural status of the institution and the programme embodying the pedagogy employed, will concomitantly loose value and credibility, since these are taken in every case as socio-culturally constructed and ever coming into being. They are the cumulative result of an ever-increasing number of individual, and essentially subjective choices upon which the accordance of value, as consensus, is always based.

[193] See in particular, Adorno's polemical work translated from the original German as *The Jargon of Authenticity.*

Adorno, (1969), *op. cit., passim.*

[194] For Adorno, the same holds true of other existentialist phenomenologists such as Lukács and Heidegger in particular.

See Adorno, (1969), *op. cit., passim*

[195] For existential phenomenologists, such is rendered possible through communicative, dialectical and linguistic relations, most often, though not exclusively,

constituting identities for mind as inner subject for itself and as both perceptive of, and partially shaped by an outer, 'objective' and cultural world with which it freely seeks dialogue.

[196] At its most reduced level of abstraction, such a dialectic allows for the conceptualisation of a physical impossibility, as *"le néant"*, defined by Sartre in his problematic English translation, in the proposition: "nothingness ***is not***; it annihilates itself".

[197] This is the Sartrean notion of *"l'existence-pour-soi"*, or "existence-for-itself". See Sartre (1946 a and b), *op. cit.*

[198] It may be recalled that this is freely-chosen and continually so, since there is nothing 'exterior', capable of determining such choices, as explained. For Sartre, 'self'-knowledge can only be attained through constructions of meaning achieved through freely-chosen acts that in socio-historical worlds, express purposive selections of concerns, continual and active focussings of attention through choice, and locations within a dialectic of relations with 'other'. Through the effects of continual modifications of socio-cultural and historical environments over time, and the absence of stasis these imply, constraints contextualising such relations are so variable as to permit effectively unlimited range in any enumeration of finely-grained options available to individuals for choice.

See pp. 70, *et sq.*

[199] This notion of 'free choice' in determining engagement with the social and natural worlds has also received much critical attention, both from Sartre himself in his later work, and from structuralists, critical theorists and neo-Marxists, such as Adorno. Notwithstanding, it remains attractive for pedagogies of language-learning and their associated forms of assessment and evaluation. Choice suggests that engagement with the social and natural worlds occurs subjectively through the process of private and selective 'focussings of attention' as explained. These may be understood as expressions of free-choice by subjects, since any stimuli available for perception and sourced in the world outside 'self', may be either freely ignored in entirety, or individually and purposefully selected from the infinite range, constantly presenting itself to conscious minds through the faculties of perception at any given point in time and in any given material context. Indeed in contradistinction, the subjective mind is also free to choose reflectively to 'turn in on itself', and focus on the features of the existence of its own 'interior life'. Such metacognition allows meaning to be attributed to organisations of thought according to individual schemata. The alternative postulations of primacy for the existence of the material, outer world that subsequently shapes mind to its own larger forms, as with Adorno, give no account of possible phenomena and capacities, realised through free selections of points of interest for focussing private attention. Hence related effects are not discussed.

See for example the modifications to existentialist phenomenology posed in the *Critique of Dialectical Reason.*

See Sartre (1960), *op. cit.*

See also Adorno, (1969), *op. cit.*

Authenticity in the World of Education

[200] 'Intersubjectivity' has been defined for example by Rogoff, as:

> "shared understanding based on a common focus of attention and some shared presuppositions that form the ground for communication".
> See Rogoff, (1990), *op. cit.*, p. 71.

The term is used in the present research in this sense.

[201] For research in assessment contexts, the effects of teaching and learning as processes in themselves, are ignored. Only the products of such processes receive scrutiny. They are recorded in performance and measured through reference to canons for any particular, named language. In this, measurements are categorised as response to task set, coherence in message, appropriateness in presentation and willingness to interact with 'others'.

[202] As declared by the **IBO** and recounted in Chapter 2.
See pp. 44, *et sq.*

[203] It is therefore not intrinsically relevant to ask whether linguistic 'culture' is established and represented at the predominantly psychological level of micro-groups of two communicating individuals, or at the more sociological level of educational classes, of schools as institutions, of larger societies, or indeed of whole language, 'national', or supra-national entities, besides any collectivity lying between such polar groupings.

However, performance in a discrete 'language', accepted as assessable by the **IBO** and categorised in *Groups 1* and *2* of the *Diploma Programme*, appears referenced to socio-cultural and linguistic 'standards', be these legally defined by a state Academy, (as in the case of French), or not. Some languages are indeed adopted as 'official' for communication within the organisation (these being respectively 'English', 'French' and 'Spanish').

It is interesting to note that the **IBO** makes no official pronouncements on how it may define the corpus of vocabulary and structures of any given 'language' accepted for assessed communication, other than in the case of *Ab Initio* programmes, designed as two-year courses for beginners.

[204] For an introduction to this view of sociocultural theory, see Lantolf, (2000), *op. cit.*

[205] McDermott for example, claims that "languages acquire their speakers". Language and culture can be related as engagements in which these categorisations are no longer represented as:

> "scripts to be acquired, as much as they are conversations in which people can participate. The question of who is learning what and how much is essentially a question of what conversations they are a part of; and this question is a subset of the more powerful question of what conversations are around to be had in a given culture".
> See McDermott, (1999), *op. cit.*, p.18.

The Concerns for Pedagogy and Learning

[206] To recapitulate, one of the key research objectives has been to categorise, describe, exemplify and analyse from empirical evidence, as well as from theoretical

literature, features distinguishing the integrative use of any language in authentic communication. In this, authentic language use is viewed as an over-arching concept, supplying a set of benchmarks for investigating task-design and the devising and application of criteria for moderated assessment and evaluation of the outcomes of pedagogy and learning, through either internal assessment, or examinations, or the combination of both.

[207] For a summary of the Piagetian approach, see for example, the work of Wood (1988), *op. cit.*

[208] See Wood (1988), *op. cit.*

[209] See for example, Wood (1988), *op. cit.*

[210] Goldfarb (2000) for instance, has summarised Vygotskian theory thus:

> "Learning is a constructivist activity. Cognitive development is a process in which language is a crucial tool for determining how the child will learn how to *think* because advanced modes of thought are transmitted to the child by means of words. "Prior to mastering his own behavior, the child begins to master his surroundings with the help of speech." Once the child realizes that everything has a name, each new object presents the child with a problem situation, and he solves the problem by naming the object. When he lacks the Word for the new object, he demands it from adults.
>
> The early word meanings thus acquired will be the embryos of concept formation. "A problem must arise that cannot be solved otherwise than through the formation of new concepts." During the course of development everything occurs twice. For example, in the learning of language, our first utterances with peers or adults are for the purpose of communication, but once mastered they become internalized and allow "inner speech." "Thought undergoes many changes as it turns into speech."
>
> See Goldfarb (2000), *op. cit.*

[211] It should be noted however, that one of the central concepts of Vygotskyan theory is the view of human mind as 'mediated'. This implies that primacy in Vygotsky's scheme, as with Adorno, is given to the existence of the external and material world whose contingency with individual mind serves as a primary feature for conditioning, and in the case of human society, for 'enculturing' all understandings and expressions of consciousness and thought. For Vygotsky, human individuals, through the use of tools and labour, and in the case of interpersonal relations using signs and language as symbolic tools and dialogic 'labour', can shape and change the world into which they are born and by which they are initially shaped.

See for example, the discussion on this aspect of Vygotsky's thought in Lantolf (2000), *op. cit.*

[212] This term is used by researchers such as Rogoff, with reference to intentional intersubjectivity as the communicative negotiation of meaning between differing subjective perspectives for reducing ambiguity in socio-linguistic environments, and in particular for establishing through prior, individual choice, jointly-negotiated foci of attention.

See Rogoff, (1999), *op. cit.*

[213] See Bruner (1986), *op.cit.*

[214] The sense here is that of the original, Latin derivation from the verb *educare*, or "to lead *outwards*".

[215] This work is Sartrean-inspired, explicitly 'revolutionary' and libertarian. Freire for example, stresses the importance to learning of "authentic dialogue" situated in creative, yet critically-aware relationships between teachers, students and the materials chosen for use in learning in communicative, collaborative and jointly dialogic "acts of knowing" that stand as "transforming acts upon the world". He defines inauthenticity in dialogue as the phenomenon of being "unable to transform reality", in which "reality" is understood as the creative, temporally ever-becoming nexus of relations that are communicatively intersubjective and contextualised within 'the world'. In Freire's perspective:

> "Education which is able to resolve the contradiction between teacher and student takes place in a situation in which both address their act of cognition to the object by which they are mediated. [.....] Authentic education is not carried on by "A" *for* "B" or by "A" *about* "B", but rather by "A" *with* "B", mediated by the world – a world which impresses and challenges both parties, giving rise to views or opinions about it. These views, impregnated with anxieties, doubts, hopes, or hopelessness, imply significant themes on the basis of which the program content of education can be built."
> See Freire (1996), *op. cit.,* Chapter 3, most notably pp. 69 –74.

[216] See Fairclough (1989), *op. cit.*

[217] See Glaserfeld (1989), *op. cit.*

The following quotation may succinctly summarise the ontological and epistemological perspective of such neo-Piagetianism:

> "Knowledge is not an iconic representation of an external environment or world, but rather a mapping of ways of acting and thinking that are *viable* in that they have proven helpful to the acting subject in attaining experiential goals. Second is the idea that this kind of knowledge is under all circumstances the result of an individual subject's constructive activity, not a commodity that somehow resides outside the knower and can be *conveyed* or *instilled* by diligent perception or linguistic communication. Third is the idea that language is not a means of transporting conceptual structures from teacher to student, but rather a means of interacting that allows the teacher here and there to constrain and thus to guide the cognitive construction of the student."

[218] See for example, the exposition of Vygotsky's thinking in this domain in Britton (1987), *op. cit.*

[219] See Bruner (1986), *op. cit.*

Indeed, in the Vygotskian rendering of epistemology, language is a product of social interaction and experience, serving in turn to structure and give direction to thinking within linguistically-constraining frameworks, thus embedding the formation of concepts within specific history and culture.

See Vygotsky (1978), *op. cit.*

[220] See for example, Bruner, (1996), *op. cit.*

[221] Bruner explains as follows:

> "Whether private meanings exist is not the point; what is important is that meanings provide a basis for cultural exchange. On this view, knowing and communicating are in their nature highly interdependent, indeed virtually inseparable. For however much the individual may seem to operate on his or her own in carrying out the quest for meanings, nobody can do it unaided by the culture's symbolic systems. It is culture that provides the tools for organising and understanding our worlds in communicable ways."

See Bruner, (1996), *op. cit.,* p. 149.

[222] See Bruner, (1996), *op. cit.*, p. 157.

Furthermore, in underlining the intersubjective possibilities for variance in any assessment or evaluation of meaning in communication, Bruner claims that:

> "a culture's judgements about the idiosyncratic construals of its members are rarely unequivocal".

That is, culture is always 'multivocal' and dependent on principles, both explicit and implicit, of what is deemed 'tolerable' in the exploration of the boundaries of both 'self' and the culture in which it is situated.

See Bruner, (1996), *op. cit.*, pp. 157 – 158.

[223] See Bruner, (1996), *op. cit.*, pp. 167 – 168.
See Bourdieu (1991).
See also Jenkins, (1992).

[224] See Bruner (1999), *op. cit.*

[225] Indeed, for Bruner, the axiological associations of choice in the construction of action by 'self' as an agent situated in the world, implying responsibility for the results of such choice, give rise to a tenet of 'self-esteem', central to the construction of valued identity and the very concept of functional selfhood.

See for example, Bruner, (1996), *op. cit.*, pp. 172 – 173.

[226] However, within situated culture, the "intentional stance of the learner", providing evidence for the "interdependence of cognition and affect", is a central feature. As reported by the **Open University**:

> "How learners *feel* about their abilities and their interest and motivation in learning particular things, fundamentally influences their engagement with tasks. This is what Bruner refers to in his self-esteem tenet."

See **Open University,** (1999), *op. cit.*, p. 63.

[227] Indeed, Bredo, in following Rorty, sees language as the tool of individual *volition* that is intentional in its direction, in that language may be viewed as:

> "strings of marks or noises which organisms use a tools for getting what they want."
> See Bredo (1999), *op. cit.,* p. 35.

[228] See Bredo (1999), *op. cit.,* pp 24 – 25.

[229] See Bredo's discussion of Rorty's critique of symbol-processing theory. Bredo (1999), *op. cit.*, pp. 28 – 29.

[230] See Bredo (1999), *op. cit.*, p.32.

Interestingly, Bredo summarises his position as follows:

> "Each [assessment activity] is the result of a dialogue, a way of relating, or mutually modulating activity in which person and environment (ideally) modify each other so as to create an integral performance. [.....] A successful person acts with the environment, shaping it to modify himself or herself, in turn, and then to shape the environment, and so on, until some end is achieved. [.....]
> The production of a well-coordinated performance involves a kind of dance between person and environment, rather than the one-way action of one on the other. "
> See Bredo (1999), *op. cit.,* pp. 33 – 34.

[231] See for example, Lave and Wenger, (1991), *op. cit.*

In some aspects, this model of learning accords with the assessment model of the **IBO**. The latter includes fully continuous and interactive forms of assessment within *Internal Assessment*. 'Masters' as teacher-guides, facilitators and interlocutors are in interaction with 'apprentices' as candidates responsible for the presentation of original work, and fellow interlocutors.

See Chapter 2, p. 44, *et sq.*

More formal, external assessment under point-in-time examination conditions equates with competence testing of the 'apprentice' under controlled conditions, for producing responses to a task in hand. In *Written Production,* imposed constraints are loosened by the availability of task-choice, choice of genre and of content in the creation of a written artefact, (albeit with a need to respect intellectual, cultural and linguistic appropriacy), and choice of maximum length of response. These are the key requirements for supplying evidence allowing the full application of all assessment descriptors.

In other words, the **IBO** scheme may be understood as granting a substantial degree of independence for the candidate to interpret test rubrics and constructs, even though such independence is not total. Authenticity as a quality under assessment is desirable for optimal performance, since the design, with its descriptors for higher levels of attainment promotes the demonstration of a clear awareness of 'self'. It seeks to favour motivated expression through language that indicates self-reflection and self-reflexivity, evidently necessary for the production of responses that combine evidence for 'personality', 'imagination' and 'convincingness'.

See **IBO** (1996), *op. cit.,* pp. 37 – 50.

See also Chapter 2, p. 39, *et sq.,* Chapter 6, *Table 6.8*, p. 217, *et sq.*

[232] See Rogoff, (1999), *op. cit.*

[233] See Rogoff, (1990), *op. cit.*, pp. 69 – 71.

[234] For Rogoff, 'guided participation' forms a perspective on learning that is not concerned with explicit pedagogy. It focuses on four key processes of all educational activity, summarising and further developing those of Bruner. Such participation may be typified as:

- an engagement of 'self', or focussing of attention by the learner on the activity in hand;
- an engagement through a joint establishment of foci for attention with 'others' who assume definable social roles in such participation;
- an engagement within a socio-temporal context and linguistic *milieu* that is typified by communal culture, a given linguistic 'norm' and established social traditions;
- an engagement for a purpose in attaining goals set either by 'self', or by 'other', or imposed by the environmental setting for activity.

See Rogoff, (1999), *op. cit.*

[235] See for example, Lave and Wenger, (1991), *op. cit.*, pp. 83 – 85.

Communicative Language Acquisition and Use

[236] Chomsky for example, conceptualises an individual's mental representation of structures in language as "competence", distinguishing them from productive use in real-life situations as "performance".

See Chomsky (1965) *op.cit.*

[237] Here it may be recalled that for the **IBO**, communication and interaction are typified by "communicative" language use, in that this focuses "principally on interaction between speakers and writers of the target language". The most significant aim of the organisation therefore promotes productive and situated use of the given language within contexts that are defined as "social", "academic" and "cultural" under the programme *Objectives*.

See: **IBO** (1996b), *Nature of the Subject: Language B,* p. 41, *et sq.*

[238] See Chomsky (1965) *op.cit.*

[239] See Hymes (1971, 1974, 1977), *op. cit.*

[240] Combining the approaches of Chomsky and Hymes, Widdowson for example categorises and distinguishes between the learner's knowledge of formal properties for discrete language structures as 'usage' and command of these structures for effective communication with others as 'use'. 'Usage' therefore permits communication that can be assessed and evaluated in 'use'.

See Widdowson (1978), *op. cit.*

[241] This for example, is usefully summarised by Orwig (1999), *op. cit.*

[242] See for example, Orwig (1999), *op. cit.*

[243] Typically, these are the aspects of communication in daily use, through routine communicative exchange for functional, or simple interpersonal and cultural purposes.

See for example, the objective broadly and imprecisely defined for the *French B, Standard Level* programme as "correct comprehension and usage of oral and written forms of the language as frequently encountered in various situations".

See **IBO** (2002b), *op. cit.,* p. 10.

[244] See Chapter 2, *passim.*

[245] Key aims of the **IBO** programme, relating to notions of communicative, authentic language use, are summarised as the promotion of the following:

- Accurate and effective communication with others through the use of the target language in speech and writing;
- Communication with others that is transactionally and socially contextualised.

See Chapter 2, pp. 39, *et sq.*

[246] From **IBO** *Paper Specific Instructions* to examination designers, as related in Chapter 2, examination tasks should serve as stimuli to authentic language production by candidates. The chosen response should be linguistically and culturally contextualised in a manner appropriate to, and thus determined by the specific design of the task. Implicitly, through covering "a range of interests", and being "relevant and interesting [for] a 17 -18 year old student", the designs should also encourage motivated responses that require expression in writing in the target language.

See **IBCA** (2001c), *op. cit.*, p. 5.

[247] Dodson (1967) for example, distinguishes between language as 'medium' and 'message' level communications. Thus for example, when students are being taught to say how old they are *("Tu as quel âge?")*, they are merely practising a given language structure solely to master the construction. Teachers probably know the age of their students, and students also know that the teacher knows their age. According to Dodson, they are all performing at 'medium' level, that is practising how to speak the language but with no added purpose.

Dodson explains a situation:

> "Suddenly, a curious member of the class raises his hand and asks the young lady teacher *"Tu as quel age?"*. This is language being used at a totally different and higher level, i.e. 'message' level (the pupil doesn't know the teacher's age, but actually uses the construction practised at the 'medium' level for a specific purpose, namely that of finding out the teacher's age!"

For Dodson, language must be rehearsed at 'medium' level before being exercised at 'message' level. The problem is that many teachers never go beyond 'medium' levels by using language for 'true' or authentic purposes of sending and receiving 'messages'. Teachers have taught students 'about' language, about its patterns and rules, rather than using it actively for 'real' purposes.

See Dodson (1967), *op cit.*

[248] Indeed, at the *Language A2* level, this feature is now made explicit in the statement under the subheading of *Classroom Environment*, that:

> "Teaching must be provided in the target language, and learning should be placed in the contexts that prepare the students for actual use of the language."
> **IBO** (2002a), *op. cit.,* p. 15.

For *Language B,* the latest requirements state that *inter alia*:

> "teachers should aim to provide a typical monolingual environment where teaching is provided in the target language and learning is placed in a context that would be familiar to speakers of that language."
> **IBO** (2002b), *op. cit.,* p. 13.

In this respect, it should be held in mind that the latest statements of the organisation represent an evolution of the programme, rather than a change of approach, and that the range of language levels offered form a single continuum, or "spectrum".

See **IBO** (2002a, 2002b), *op. cit.*

In this context, statements relevant to the *A2* programme, whilst not the focus of research, represent perhaps the most unambiguous **IBO** declarations of its conceptualisation and use of 'authenticity' as a key notion, colouring more ambiguous usage at 'lower' levels, such as those for *Language B*. Indeed the border between programmes is intentionally ambiguous, with teachers exhorted to place students "appropriately" to represent an "adequate challenge" for learning, and avoiding the "amass[ing of] points in an educationally sterile fashion".

See **IBO** (2002a, 2002b), *op. cit.*

[249] It may be recalled that for the **IBO**, the 'communicative classroom' should provide opportunities for rehearsal of real-life situations and provide opportunity for real communication. The emphasis is often on creative role-plays, simulations, surveys, projects, short scenes and so forth. All are considered as favouring the spontaneous, sometimes improvised production of authentic language.

It may also be noted that for assessment purposes for *Group 2 Languages, Language B* at both levels, *Social Objectives* are commonly defined as a demonstration of the ability "to respond to the complex demands of day-to-day communication". The recognition of implicit meaning and attitude is isolated as a requirement for *Higher Level* assessment only. Together, they relate to the aims of transactionally and socially-contextualised communication with others, in that the demands of the programme are categorised as:

- "obtaining information from written and oral sources;
- processing and evaluating information from written and oral sources;
- communicating or corresponding with users of the target language in both formal and informal situations;
- making social or professional contacts with people who live and work in the country or countries concerned;
- expressing views and opinions on issues if general interest;
- expressing feelings."

See **IBO** (1996b), *op. cit.,* pp. 9 – 10.

[250] This may be compared with the **IBO**'s approach recommending that target languages be taught through the exposure of students to "a wide range of oral and written texts of different styles and registers", with recourse to "authentic materials [.....] wherever possible", and maximum use of this language.

See **IBO** (1996b), *op. cit.,* p. 6.

[251] See Krashen, (1981), *op. cit.,* pp. 6 – 7.

The Identification of Components of Authentic Language Use

[252] See Edwards and Mercer (1987), *op. cit.*

[253] See Van Lier (1996), *op. cit.,* especially Chapter 6.

[254] This concept is adapted by Van Lier from the work of Charles Peirce. Put briefly, it refers to a sequencing of concepts in families of three, whereby each individual unit of a triad, whilst occurring in an hierarchical order of "firstness, secondness and thirdness", constitutes the whole through the workings of fully interdependent relationships with the others. These cannot therefore be completely understood by isolating any one unit from the others.

See Van Lier (1996), *op. cit., passim.*

[255] This includes, it is supposed, any given assessment model associated with such curricula, although assessment activity lies in a domain that is undeveloped by Van Lier.

[256] Van Lier, (1996), *op. cit.,* pp. 133 – 134.

[257] In Chapter 5, it will be seen that these assessments may serve as a source of triangulating data for evaluating the meaningfulness and coherence of the **IBO**'s own criteria and procedures for the assessment, moderation and evaluation of exemplars of language, as produced under the published rubrics of both internal assessment and external examination.

[258] Van Lier, (1996), *op. cit.,* pp. 135 - 136.

[259] These are features employed as discriminators in the assessment grid developed as a key research instrument, described in Chapter 5 and illustrated in **Appendix 3.**

[260] See Van Lier, (1996), *op. cit.,* Chapter 6.
See Csikszentmihalyi (1990), *op. cit.*

[261] Such 'authentication' is a process defined by Van Lier as follows:

> "It establishes *relevance*, and it *endorses, rejects,* or *revises* prior utterances. Inauthentic discourse then happens when defectiveness (e.g. a discrepancy of interpretations) occurs which is not (successfully) repaired. [.....] In many cases it may be relevant to use Habermas' concept of *systematically distorted communication*, which

> can be described as the result of 'a confusion between actions oriented to reaching understanding and actions oriented to success' (in which case a situation of unconscious deception obtains); Habermas 1984: 332"

. Furthermore, Van Lier defines authenticity as:

> " the result of a process of authentication, a validation of classroom events and language, and an endorsement of the *relevance* of the things said and done, and of the *ways* in which they are said and done."

He comments that:

> "such authenticity results from self-determination (knowing-what-you-are-doing), a commitment to understanding and to purpose and transparency in interaction. As Sartre says in *Being and Nothingness*, like individuality, such authenticity is not given, it has to be *earned* (1957:246)"

(Van Lier's italicisations throughout).

See Van Lier, (1996), *op. cit.,* Chapter 6, pp. 127, 133 and 128 respectively.

[262] See Chapter 5.

Authenticity and the Measurement of Linguistic Attainment

[263] The form of these tasks may be set as presentations, interviews, tests and examinations, in both internal and external assessment exercises for 'second' and 'foreign' languages, for reporting to parties external to the processes involved. This is the context of the empirical study completed.

McDermott provides a useful introduction in a discussion of "*The Continuum of Arbitrary Demands and Left-Out Participants*", which focuses on an exemplary 'problem' student named Adam. McDermott explains thus:

> "In everyday life, Adam can use any resources to get a job done. [.....] School tasks are different from this in that a person is often restricted in what he **[sic]** can make use of; procedure is of the essence. On tests, this trend is exaggerated. What else is a test but an occasion on which you cannot use any of the resources normally available for solving some problem [.....] Is it possible that Adam is better understood as a child who is faced not by increasingly more difficult tasks, but increasingly more arbitrary tasks? [.....] At the very least, cross-cultural psychology has been extraordinarily clear in showing how the various kinds of smartness could be reduced to apparent ignorance in the face of culturally arbitrary and cross-culturally foolish tasks [.....]
> Could Adam be disabled on his own? Only if he could work on a task that was not culturally defined and had no consequences for his life

> with other; that not being a possibility, he can only be disabled through his interactions with others"
> See McDermott (1999), *op. cit.*, pp 10 – 15.

Whilst the discussion appears to assume prior acquisition of a common language in which Adam's 'conversations' with others may take place, there is no intrinsic reason why such should not remain true in the learning of others through any particular language, once a minimal starting point has been established in mutual choice to communicate together through such language. It is in this sense that:

> "Context is defined not as something 'into which someone is put, but an order of behaviour of which one is part'."
> See **Open University** (1999), *op. cit.*, p. 47.

[264] McNamara (2000) for example, defines such 'high-stakes testing' as:

> "Tests which provide information on the basis of which significant decisions are made about candidates, e.g. admission to courses of study, or work settings." (*op. cit.*, p. 133)

CHAPTER FOUR: The Literature of Assessment

The Design and Standardisation of Communicative and Assessable Tasks

[265] In this instance, this review ignores **IBO** categorisation of 'levels' of language by discrete group, as *A1, A2, B* or *Ab Initio*, as well as by *Higher* or *Standard Levels*, in all but the latter group. It may be recalled that by **IBO** definition, *A1* represents a course of literary study for students in their 'native', or 'best' language, or alternatively in the working language of the bulk of their educational experience; *A2* represents a course in language-handling and textual analysis for bilingual, or highly experienced 'second' language students capable of using the language in all working situations; *B* represents a course in language-handling, including the optional study of literature, for learners exposed to the language as a 'foreign' language, yet predominantly working in other languages; *Ab Initio* represents a course of language-acquisition for complete, or near beginners, covering the rudiments of structure and lexis essential for everyday, communicative working purposes. In theory, the coherence of this range of programme offerings is to be viewed as a continuum of powers in comprehension and expression from *Ab Initio* to *A2*, as compulsory *Group 2 Languages*; from beginner, teacher and course-dependent 'apprenticeship' to rich and confident, wholly independent 'mastery'.

[266] McNamara, (2000), *op. cit.*, p. 13.

[268] The work of Lado outlines the major features of this school of though in the context of language testing.
See Lado (1961), *op. cit.*

[269] See Garnham (1985), *op. cit.*, Chapters 1 and 9.

Psychometrics and the Approach of Psycholinguistics

[270] See Garnham (1985), *op. cit.*, Chapter 1.

[271] More often than not, such knowledge and skills are evaluated as quantitatively equal. To cite one influential example in the case of the English and Welsh *GCSE* examination, 25% of total scores for any comprehensive assessment is allocated discretely to each of the four skills and aggregated to 100%, but with high levels of attainment in writing a necessary pre-condition for the award of the highest grades.

With the **IBO** *French Language B* programme, whilst statements defining *The Nature of the Subject* and the *Syllabus Outline* emphasise equality in weighting of the same four skills, for a variety of reasons, the assessment and evaluation practice does not.

See **IBO** (1996b), *op. cit.,* pp. 6 and 11, and the descriptions of the programme outlined in Chapter 2.

[272] See Garnham (1985), *op. cit.*, Chapters 1 and 9.

With psychometrics, a commonly-found assessment activity requires candidates to select 'correct' answers, without recourse to 'self-expression' through language. Typically, this takes place in multi-choice, image-matching, pairing, and other forms of discrete-point testing.

Communicative Language Use in Assessment and Evaluation

[273] See Oller, (1979), *op. cit., passim.*

[274] See McNamara (2000), *op. cit.,* p. 15.

[275] Examples for close-testing examination are the elimination of fixed categories of linguistic elements (such as prepositions or verb endings), or of frequency-based features of written expression (such as those provided by every *n*th. element in reading-text reconstruction and completion).

[276] See for example, Hymes (1974), *op. cit.*

[277] See Canale and Swain, (1980), *op. cit.*

[278] Such organising categorisations may usefully be compared with those developed by influential structural linguists such as Halliday and Hasan. With these analysts for example, all communicative language may be categorised through reference to Halliday's concepts of *linguistic field, tenor* and *mode*. In this respect (and with a certain amount of injustice to the sophistication of Halliday's thought, through over-simplification), *field* may be taken as broadly synonymous with the categories of *strategic* and *discourse competence*, *tenor* as broadly synonymous with *sociolinguistic competence,* and *mode* with *grammatical competence*.

See Halliday, (1975), *op. cit.*

To this, Hasan has added the concept of *texture*, achieved through the use of linguistic, *cohesive ties* as a key feature in the structural and analytical conceptualisation of communicative quality and value in 'meaningful' language production.

See Hasan, (1989), *op. cit.*

[279] See Bachman (1990), *op. cit., passim.*

[280] See McNamara, (2000), *op. cit.,* pp. 20 – 21.

Criterion-Referencing for Measuring Language Use

[281] See the *Guide to the Programme: Language B*, where it states for example:

> "The method of assessment used by the **International Baccalaureate Organisation** (**IBO**) is criterion-referenced, not norm-referenced. That is to say, the method of assessment judges candidates by their performance in relation to identified assessment criteria and not in relation to the rest of the candidates"
> (**IBO** (1996b), *op. cit.,* p. 34.

In this programme, final evaluation requires synchronic, point-in-time, language production for internal assessment and external examination, rather than diachronic, portfolio, or records of achievement-based collections of evidence.

However, as reported in Chapter 2, samples of oral production for *Internal Assessment* are collected at various points in time over the course of the final year of preparation for assessment and evaluation by the **IBO**.

In this context, reference may usefully be made to McNamara's (2000) monograph devoted to issues of language testing.

See McNamara, (2000), *op. cit., passim.*

[282] See McNamara, (2000), *op. cit.*, p. 132.

[283] See McNamara, (2000), *op. cit.*, p. 132.

[284] See Gipps (1994), *op cit.*, especially Chapters 5 and 6.

[285] See Glaser (1963), *op. cit.,* p. 520.

[286] See Gipps (1994), *op. cit.*, p. 98.

[287] Further discussion of these points follows later in this chapter.

[288] See Gipps (1994), *op. cit.*, p. 98.

[289] See Gipps (1994), *op. cit.*, p. 98.

[290] See Meyer (1992), as reported in Gipps, (1994), *op. cit.*, p. 99.

[291] Evidently, given the design of the relevant **IBO** programmes as the subject of the research, such distinctions are significant.

[292] However, for reasons of reliability in drawing comparisons across differing teacher-assessors, final evaluation is largely dependent upon the assessment of performance in a set presentation and interview, common in format to all candidates and assessed by an independent *Internal Assessment Moderator,* as described in Chapter 2.

[293] It should be remembered that this term is defined by McNamara, as:

> "The area of knowledge or skill, or the set of tasks constituting *criterion* [author's emphasis] performance, and which is the target of the test."
> (McNamara (2000), *op. cit.*, p. 233.)

In the context of **IBO** programmes, these are described and summarised in Chapter 2.
See Chapter 2, p. 55, *et sq.*

The tripartite classification of language use in this way reflects structural categories for language description and analysis, developed by influential linguists such as Halliday (1975)

Thus, the **IBO**'s categorisation of criterion descriptors for *Task and Message* may be seen to relate to Halliday's conceptualisation of communicative structures in discrete elements, generally categorisable as linguistic *field.* The conceptualisation and categorisation of essential elements of oral *interaction* on the part of the producer of language at least, may reflect Halliday's concerns in investigating the structural concept of linguistic *tenor*, further modified and partially re-categorised by associates such as Hasan (1989) who emphasises the significance of *texture* in the production of meaningful text, through the 'appropriate' use of *cohesive ties. Language*, as the third category established by the **IBO** for assessment purposes, may be taken to relate to the broadly-based, Hallidayan concept of linguistic *mode.*
See Halliday (1975), *op. cit.*
See also Hasan (1989), *op. cit.*

[294] See Note No. 293 above.

[295] See Chapter 2, p. 55, *et sq.*

[296] See Chapter 2, p. 55, *et sq.*

It may be noted however, that overlap between categorisations evidently occurs, in that in one example, *fluency* may be seen as a consideration for assessment of oral *Interaction,* written *Presentation* and *Language,* whether oral or written.

[297] It is perhaps noteworthy that in further breakdowns of criteria by detailed descriptor, neither explicit and discrete evaluation by point-value, nor weighting by discrete categories for assessment is given.
See Chapter 2, p. 55, *et sq.*
See also: *Guide to the Programme: Language B*, **IBO** (1996), *op. cit.*, pp. 38 – 50.

[298] Notably, these refer to issues of material and financial resourcing in the production and administration of assessments from many, differing centres; the provision of appropriately-trained raters; and in certain cases, the requirement to report procedures and data to higher, administrative authorities, be they at local, regional,

national or even international levels. Distance, and the need for varied means of communication between test administrators and test centres, create questions for test security.

See McNamara (2000), *op. cit.*, Chapter 3.

[299] See McNamara (2000), *op. cit.,* p. 25.

[300] See McNamara (2000), *op. cit.,* pp. 25 - 26.

[301] See McNamara (2000), *op. cit.,* p. 27.

[302] See for example, Bourdieu, (1991), *op. cit.*

[303] See for example, Fairclough, (1989), *op. cit.*

[304] Furthermore, it should be recalled that the **IBO**'s *a priori* categorisation of its programmes, defined by levels of language acquisition, is deemed successively: "for highly competent speakers of the target language" (for the *A2 Programme*); "foreign" but for those with "previous experience of learning the language" (for the *B Programme*); and as "foreign" for "beginners" (for the *Ab Initio Programme*).

Indeed, such categorisation creates significant further constraints that may have evident impact through 'washback', in determining the aims, assessable objectives, modalities, methods and content of teaching and learning.

See **IBO** (1996b), *op. cit.*, p. 3.

See Chapter 2, p. 36, *et sq.*

[305] In this context, it is appropriate to recall that the **IBO** exhorts teachers and institutional programme co-ordinators to 'guide' student choice, though the organisation does not prescribe any possible choice as a pre-requisite for registration in a particular programme of assessment. Accordingly it is stated that:

> "Teachers and **IB** coordinators should ensure that, as far as possible, students are following the course which is most suited to their needs and which will provide them with an appropriate academic challenge".
> See **IBO** (1996b), *op. cit.,* p. 5.

Further analysis and discussion of the problems that this raises is included in Chapter 6.

[306] See Sanderson (1997), *op. cit.*

[307] See Sanderson (1997), *op. cit.*, p. 77.

[308] See Bachman and Cohen (1998), *op. cit.*

[309] See Bachman and Cohen (1998), *op. cit.*, Chapter 1.

[310] See Bachman and Cohen (1998), *op. cit.*, pp. 2 – 3.

[311] See Bachman and Cohen (1998), *op. cit.*, pp. 2 – 3.

[312] See Bachman and Cohen (1998), *op. cit.*, pp. 22 – 23.

The Standardisation of Examination Tasks

[313] That is, in reiteration of the exemplary cautioning of McNamara:

> "Defining the test construct involves being clear about what knowledge of language consists of, and how that knowledge is deployed in actual performance (language use). Understanding what view the test takes of language use in the criterion is necessary for determining the link between test and criterion in performance testing."
> McNamara, (2000), op. cit., p. 13.

[314] These are traditionally taken in conformity with immutable, known rules of stable language, socio-politically 'approved' as a standard.

[315] In exemplifying the phenomenon of 'familiarisation' insofar as **IBO** moderation and evaluation procedures are concerned, it was noted however, that the longitudinal records of relevant *Subject Reports* are not consulted during respective *Grade Award Meetings*. Yet one such report, as shall be seen later, devoted to the analysis of latest candidate work, with associated comments and recommendations for future work for *French Language B*, makes significant reference to the effects of 'familiarisation', or 'washback'.

Such effects form an important aspect for consideration, as signalled in the literature of assessment. They are further described and discussed in Chapter 6, particularly with respect to the moderation of *Internal Assessment* for the May 2001 examining session in *French Language B, Standard Level.*

The concept and effects of 'familiarisation' also formed a significant element in focusing research attention for data-collection from the **IBCA** *Grade Award Meeting* for *German Language B* for the May 2001 session, reported in Chapter 2 and detailed in **Appendix 2**.

[316] In developing a framework of levels for the comparison of language tests, the **Council of Europe** has developed a *Common European Framework of Reference for Language Learning and Teaching.* This comprises six main levels, specified as *A2* and *B1* and labelled respectively as *Waystage User* and *Threshold User* levels.

From 1971, the Council established divisions of language-learning in a hierarchical order of levels, each of which could be credited in assessments. Within a communicative approach, the Council stressed the necessity of basing curricula on perceived learner 'needs' rather than atomised language structures. A major outcome was the specification of a *Threshold Level* by Van Ek (1975), proposing a communicative model for the description of language knowledge and skill. A lower level specification was also produced, under the name *Waystage Level.* (In collaboration with Trim in 1991, Van Ek revised and updated versions of both levels, published as *Threshold Level 1990* and *Waystage Level 1990*).

In 1996, Van Ek and Trim further developed this hierarchy of level description, for the Council, with an additional level, known as *Vantage.* This retained the existing structures and constructs set for earlier descriptions, so as to establish coherent progression for learners, providing general objectives intended cognitively and linguistically to be as far above *Threshold Levels* as *Waystage Levels* are below them.

Vantage Level goes beyond the minimal means needed by learners to transact the business of everyday life and to make social contact with those encountered in another country. In linguistic terms, range in grammar and vocabulary is extended, as is the demand for greater control of sociolinguistic and discourse strategies, together with the display of greater sociocultural awareness. This should permit learners to develop flexibility in dealing with the unexpected and with complexities in daily living, including use of the target, second language in work, or for study purposes.

See Van Ek, (1975), Van Ek and Trim, (1991, 1996), *op. cit.*

[317] In conformity with this philosophical intent and aim for commonality, and as shown in Chapter 5, a single system for the measurement of features of linguistic authenticity at any level and in any language has been devised for the research.

This has led to the development of the instrument designed to unify theory and practice in the assessment and evaluation of significant evidence of any recorded example of task-based language use, described in Chapter 5 and illustrated in **Appendix 3**.

[318] The unification of constructs for task-design, a foundation for researching the validity of the **IBO**'s language assessment and evaluation systems is revealed in statements describing the organisation's programmes, as reported in Chapter 2. In common they respectively cover an overall, epistemological and ethical philosophy, with consonances in both the published aims for each of the three groupings of *Group 2 Languages*, and the specification of many of the graduated objectives for each level within any of these three groupings, whether *A2, B* or *Ab Initio* languages.

See *Subject Guides,* **IBO** (1996b), *op. cit.*

See also Chapter 2, p. 36, *et sq.*

As a result, the investigation of conceptualisations and usage, developed by the **IBO** and applied to groupings and levels outside the specific parameters under research, throws light on the understanding of authenticity within the bounds of *Group 2 Languages: French Language B, Standard Level.* Such research is included in the discussion of evidence, presented in Chapter 6.

<u>Categorisations of Language-Performance for Evaluation Purposes</u>

[319] See McNamara (1996), *op. cit.,* p. 12.

Administratively-speaking, **IBO** criteria for designing and standardising tasks appear to have been determined separately from those for assessing and evaluating language produced in response, though not completely so.

Rationales concern issues of establishing credibility and earning 'recognition' for programmes from external, validating institutions such as universities, as well as issues of equitable commonality and consistency in task and response requirements, across differing administrations of a single programme. Implicitly however, linkages between task-design, standardisation, assessment and evaluation result from the choices of a small group of designers and standardisers. They illustrate professional, though inevitably also individual, and possibly therefore variable understandings of linguistic 'appropriacy', and are applied differentially at both *Higher* and *Standard Levels,* across a given language grouping. The context for this aspect of programme design has been described and discussed previously.

See Chapter 2, p. 45, *et sq.*

[320] See Bachman and Cohen (1998), *op. cit.*

[321] See for example, Gipps (1994), *op. cit.*

[322] In this context, the **IBO** publishes its commitment to transparency in all its assessment and evaluation criteria and procedure, urging familiarity with these publications and their practical outcomes through regular attendance of teachers in dedicated training-sessions, held throughout the world.

See for example **IBO** (1996 a,b, 1997a, 2001e), *op. cit.*

[323] Further investigation requires research into teacher understandings, approaches and recommendations in preparing candidates for 'high stakes' assessment and evaluation, student attitudes and approaches to task-choice for formal assessment, the preparation and composition of discrete responses, whether oral or written, and the reading, checking and editing of outcomes in *Written Production*, as may be established from survey. Such extensions to research have not been completed for inclusion within the body of this project.

The Role of Examiner Training and Moderation

[324] See Gipps (1994), *op. cit.,* p. 105.

[325] All are indeed, evident features for consideration in the existing design applied by the **IBO**.

[326] See Chapter 2, p. 44, *et sq.*

[327] Given emphasis on the identification, collection, recording, description and analysis of very specifically contextualised, empirical data, and in accord with Gipps' conclusions, formal development of the theoretical considerations outlined has not been deemed central to research purposes as description, analysis and critique of an existing assessment and evaluation programme. Investigation of the effects of 'familiarisation' and 'washback' for any given model and design through training all the actors concerned have therefore, largely been put aside.

Nor have the bounds of the project permitted significant space for discussion of theoretical issues of inter-rater reliability, though the phenomena observed in **IBO** assessment, moderation and evaluation practice are indeed measured, analysed and discussed in Chapter 6. Consultation of further literature has therefore been selective.

See however, Black (1998), Gipps (1994), McNamara (1996; 2000), *op. cit.*

Grade Awards and the Relating of Scores to *General Grade Descriptors*

[328] See Gipps, (1994), *op. cit.,* p. 93.

[329] See **IBO** (1996a), *op. cit.*

[330] These are published in *Subject Reports* for each examination session.

PART III

CHAPTER FIVE: Rationales and Methods

Preface

[331] See Israel (2000), *op. cit.*

[332] These statements concern the organisation's overall philosophy in its programmes, their general aims and assessable objectives, the design and standardisation of appropriate tasks, materials and criteria for assessment purposes, together with both the procedures and criteria for the formulation of examinations, their standardisation, assessment, moderation and evaluation.

The notion was also explored for describing and analysing the shaping of language-production whether through teacher and candidate perception, attitude, approach and positioning, or through those of other 'clients' of a given programme. However, for reasons of practicability in delimiting the bounds of a single project, this research was limited and excluded from detailed analysis, reporting and discussion in the present thesis.

The Scope of Empirical Research

[333] By numbers registering, the other comparable languages are English and Spanish.

[334] However, in investigating **IBCA** moderation and grade-awarding procedures, reference is also made to German, though solely within the framework of a limited exercise for establishing validity and reliability of observationally-sourced data, through comparing procedure and outcomes in observations for French.

It should be noted that no formal investigation was made of English as a 'foreign' *Language B*, since the **IBO** recognises the situation of this subject as anomalous. In this, the candidature is considered exceptional, and therefore unrepresentative, since many candidates are registered for *English B* when the use of the language either forms a significant element of their personal background, or is the language of instruction of their school. Such candidates are in general, more appropriately described by the rubrics for *English, Language A2.*

The tendency noted was confirmed to the researcher in informal discussion with the **IBCA** *Director of Assessment* in August 2002.

[335] Further investigation of the problematic nature of such a conceptualisation of language acquisition and use has not been possible, for practical reasons. Given the design of the **IBO**'s language programmes as common to all languages, and defined as such, it has not been possible to compare evidence available from a range of other languages.

It would be desirable in future research to consider language production data, derived from **IBO** programmes and applied to non-Western, non Indo-European languages.

[336] Inevitably, though perhaps regrettably, such restriction implies further, concomitant limitation in potential for developing sophisticated understandings. In this, discrete categorisations of language, either by programme (such as the **IBO**'s *A2, B* and *Ab Initio* designations), or by level (as *Higher* and *Standard*) seem relevant in their effects and worthy of further research.

[337] This language production is often implicitly referenced by the **IBO** to norms defined for contemporary French, as generally established (but again, in significant respects not exclusively so) by the **Académie Française**, and as a second, or foreign *Language B*: a further restriction of research scope.

As related in Chapter 2, forms of French, not sanctioned by the *French Academy*, may be acceptable for **IBO** use. The most obvious and commonly encountered examples are those provided by Belgian, Canadian, Swiss and other usages attributable to 'accepted' regional variants, dialect and patois. Recognition of the expectations of an appropriate audience or readership allow for variance, as assessable in particular, under *Criterion B* of the relevant criteria: *Interaction*, in the case of language use in *Internal Assessment*; *Presentation* in the case of *Written Production.*

[338] Indeed, the production of language for measuring comprehension introduces further variables that overly complicate the research perspective, were they included as data. Traditional forms of comprehension testing, particularly of discrete lexical items and grammatical structures are either dialogically non-interactive (such as choosing correct alternatives in multiple-choice assessment), or may be distorted, not through inadequacy in comprehension, but through failure to produce appropriate language indicating successful comprehension. Hence further controls of the variables involved are required for establishing validity and reliability to a plausible degree.

In this way, and in accordance with the parameters determined for the research at its outset, the assessment and evaluation of reading comprehension and appreciation, as 'tested' in *Paper 1: Text-Handling* of the relevant programme does not form a part of the brief. The assessment of listening skills is integrated with oral assessment in the *Internal Assessment* component of the programme, in accordance with its design and rubrics.

[339] These students remain anonymous throughout the research, with personal identifications irrelevant to its nature and outcomes.

[340] For theorising language-based, communicative authenticity, they also serve to focus research attention on situated usage as a whole, rather than on comprehension as received knowledge and non-interactive skill. Hence the research moves away from the traditions of psycholinguistics and structural linguistics in assessment and evaluation, in a shift towards greater consideration of the relevance and utility of existential phenomenology and sociolinguistics for measuring authentically-produced language use, as has been explained.

In this discussion however, the identification, selection, collection, recording, description and analysis of relevant, empirical data, derived from specific *Internal Assessment* and external examination sessions, is of primary concern.

[341] Two significant issues for assessment highlight 'standardisation' and 'familiarisation' effects across equivalent assessment administrations, when measured longitudinally over time. These have direct bearing on questions of authenticity, encouraging candidates to enhance their chances of 'success' through practising and memorising the application of minimally-situated *formulae* for communication in set

responses. Such may be deemed appropriate and thereby permit the rewarding of appropriate language production with higher grades. The temptation of 'question-spotting' appears to favour inauthentic approaches to language use, through minimising the communicative possibilities of adaptation to task and situation, and ignoring the authenticity required for representing the individualised and variable concerns of 'self', in relation to the similarly unpredictable, temporally-evolving concerns of 'other'. Pre-prepared responses to anticipated tasks may well divorce them from any specific sociolinguistic, and socio-cultural context, at any given point in time.

See Chapter 4.

[342] Less formal and detailed analysis arising from experience in examining from the inception of the programme in examinations in 1996 to 2003 was also completed, further influencing the perspectives and understandings developed in the research.

[343] This professional experience was developed in employment, prior to commencing research, from the inception of the current programme for *French Language B*, introduced by the **IBO** in May 1994, with first examinations in May 1996.

A certain amount of data for *Paper 1,Text-Handling*, for *French* and *German Language B* was also collected, for reasons previously given.

[344] Together with review of existing data on the criteria and procedures structuring *Grade Award Meetings*, as observed at **IBCA** for moderating and evaluating examination work (reported in the case of *French* and *German* examining sessions in December 2000 and June 2001 respectively), these are presented in Chapter 2, and further discussed in Chapter 6.

Within this framework, standardisation is problematic in a further dimension, inherently requiring the discrete categorisation of language, to produce 'standards' that are coherent within a single subject domain and level. As a procedure, standardisation assumes concepts of 'stability' and 'complexity' at differentiated and pre-determined 'levels' of language production, as points of reference for the processes involved.

[345] This appears so for all *Group 2 Languages* programmes, whether *A2, B* or *Ab Initio*, with further constraints included by the subdivision of *Languages A2* and *B* into *Higher* and *Standard Levels.*

In the context of standardisation, these subdivisions are categorised through the specific parameters of *Group 2 Languages.* They are however undefined in the published statements for describing the *Nature of the Subject*, its *Aims, Objectives* and the *Syllabus Outlines.*

See **IBO** (1996b), *op. cit.,* pp. 6 – 23.

The Selection Of Sources of Data

Excepting publications, all **IBO** documentation is internally archived at **IBCA**.

As recounted in Chapter 2, major sources of data were provided by documents for internal use, relating to the design and administration of examinations, their assessment, moderation and evaluation. Account has also been given of the observation and reporting of **IBO** *Grade Award Meetings*, to which the researcher was invited. Attendance at these meetings was not only for research purposes, but also in fulfilment of the **IBO** policy of ensuring procedural transparency through the presence of an independent, but interested and **IBO**-remunerated, *Teacher-Observer.*

It may further be noted that the researcher was employed in this capacity for the observation of the *Grade Award Meeting* for *French: Language B* of December 2000, but in the case of the similar meeting for *German: Language B*, of June 2001, was unremunerated, and hence entirely independent.

[347] Similarly, the methodology of criterion-referenced moderation as a form of triangulation for arriving at consensual evaluations was not in itself researched, being an *a priori* given, in framing research within a set, known programme.

[348] Greater scope for data-collection, including samplings of assessments and moderations produced by others than the researcher, whilst complex, could prove significant in analysis. Were they gathered, they may improve the validity and reliability of research, thus facilitating acceptance of the plausibility of its general conclusions.

[349] This simplicity renders evidently more credible, assumptions of longitudinal stability in the interpretations and assessment, or moderation judgements of the sole rater. The significant variables of inter-rater reliability, requiring investigation in complex, multivariate analysis, are thus removed from the design. The methods for sourcing data-collection permit plausibility in drawing conclusions from analyses completed in this way.

Indeed, as will be understood from the subsequent chapter, the reliability of the researcher as an assessor and moderator has been validated by the **IBO** through its own internal procedures.

[350] The **IBO**'s methods for establishing assessment interpretations and evaluations are indeed founded on moderation, a system for triangulating assessor perspectives, in order to produce consensus in value-judgements.

[351] The moderation procedure is outlined in Chapter 2 and the resultant evidence is presented, analysed and discussed in Chapter 6.

This selection of data from candidate recordings and scripts includes complete samples, allocated by **IBCA** to the researcher as *Internal Assessment Moderator* for the May 2001, 2002 and 2003 examining sessions, and as *Assistant Examiner* for *Written Production*, from May 1996.

From the initiation of formal research, the quantitative analysis of samples of writing was mainly restricted to texts from the May 2001 and 2002 examining sessions. However, familiarity with written productions for the May sessions from 1996 to 2000 and in 2003 has also influenced understandings to varying degrees.

Copies of candidate scripts reviewed at *Grade Award Meetings*, at which the researcher as *Teacher-Observer* was present should be added to this body of data.

[352] These are described and analysed in Chapter 6.

[353] In the case of the present researcher, these statistics are given in *Note No. 388*.

Material Excluded from Investigation

[354] By this method, it was planned better to understand the procedures of individual choice by conscious 'selves', and the constraints restricting the operations of such choice, when precisely situated both as process and as product. Such factors are

central to conceptualisations of authenticity. 'Self', as initiator in producing communicative language, as interlocutor in speaking, or as a pole in the dynamics of communication relating reader and responsive writer, could thereby be better delineated. The rationales, content, and forms of the relevant mental operations of candidates under 'high stakes' examination could be investigated in context, better to establish validity within the research domain.

355 This procedure was agreed by negotiation at the outset of research, in part for ethical reasons and in part to aid the identification of relevant sources of unpublished data.

The Description and Experimental Analysis of Data

356 To recapitulate, these were first, the comparative assessment of samples of language produced orally for *Internal Assessment*, and in writing for *Paper 2* for *French Language B, Standard Level*, then the observation and recording of the proceedings of **IBCA** *Grade Award Meetings* for the moderation and evaluation of examination scripts, and finally the consultation of a sample of documents, both formal and informal, sourced at **IBCA** and intended for the internal use of the organisation in its devising and administration of examinations.

The Measurement of Authentic Language Use

357 Throughout the research, the design of this instrument was continually refined, though fundamental conceptualisations and purposes remained constant. Indeed, such development improved the interlinking of communicative task-design with assessments of authentic responses, as summarised in the review of literature.

358 Although with criterion-referencing exercises such as these, greater validity and reliability in interpretation is attainable through analysing common data, produced by a number of independent raters replicating procedures and using identical instruments, such method would have extended the research beyond feasible bounds.

It would be useful for future work in this area to compare experimental applications of such qualitative judgements with the quantifications in score that emerge.

359 To state as much obviates no claim that interpretativist approaches, fundamental to criterion-referenced assessment, may safely ignore uncontrolled aspects of method. As in all exercises that match holistic experience of real-time listenings and readings to written descriptions of 'typical' performance levels, even if discretely-categorised in similarly holistic sub-divisions of *Task, Message, Interaction* or *Presentation* and *Language*, judgements always remain to a certain degree, both imprecise and contestable. They attempt to measure the socio-cultural, temporal and existential intangibles of individual relationships between speakers and listeners, writers and readers. For establishing validity and reliability they require repeated moderation, and therein open ways to further interpretation and contestation.

In 'typical' cases, these alternatives occur with ever-decreasing frequency, intensity and variability. Indeed, assessor subjectivity is explicitly recognised in the design of the **IBO**'s assessment criteria and procedures, with a two-point variance for scoring in each descriptor category provided for the purpose.

The effects of this provision are analysed and discussed in Chapter 6.

It may be noted that positivistic evaluation, for improved triangulation of results through supplementary comparison would have required the use of sets of norm-referenced criteria that do not exist within the context of the **IBO** programme researched, would have required specific design for adaptation to the *Tasks* and *Responses* to be measured, and hence have been excluded from study.

The Measurement of Authentic Language Use

[360] See Van Lier, (1996), *op. cit.,* Chapter 6.

[361] See Csikszentmihalyi (1990), *op. cit.*

The Design of the Research Instrument

[362] The categorisations are summarised and organised into an assessment grid, reproduced in **Appendix 3**.

[363] See Chapter 2, p. 59, *et sq.*

[364] Indeed as has been commented, the inclusion of such interpretativist subjectivity within assessment procedures is seen as necessary in any authentic and meaningful, linguistic interaction between speaker and listener, writer and reader.

[365] Issues of weighting in aggregating the scores obtained for each discrete component remain however, largely unaddressed by the research.

[366] See for example, the evidence presented in Chapter 6.

[367] The same appears to hold true in the case of the researcher as *Assistant Examiner* for written production.

See Chapter 6.

[368] However, as described and discussed in Chapter 2, these include elements of choice.

In this case, all tasks have been aggregated as equal in value, without differentiation.

[369] This tendency is also noted and recorded by **IBCA** from archives of further moderations of researcher assessments under **IBO** rubrics and criteria.

The evidence is personally reported on an annual basis to each *Internal Assessment Moderator* and *Assistant Examiner*, and is recorded in the researcher's professional record, held by the researcher and **IBCA**.

[370] For this reason, the scores attributed have not been modified to equalise totalisations under all systems at a maximum of thirty points, thus obscuring the illustration of the phenomenon. As shown, the use of the enhanced model, "with

plussages", tends better to discriminate the qualities of performances above the mean, than the deficiencies of those below.

[371] Commonality of presentation is almost impossible to define for *Internal Assessment*, given the range of individual choices.

These are illustrated in **Appendix 4**, with findings discussed in Chapter 6.

[372] These anomalies are analysed and discussed in the subsequent chapter.

The Research Instrument in Use

[373] This has been defined as "recognitions of 'other' as listener or reader, and as focussing attention and linguistic interaction through respect for commonly-acquired social traditions and communicative convention, thereby allowing coherent initiations and continuations of communication".

See Chapter 5, p. 136.

[374] The results are shown in the subsequent chapter, where experimental evaluations are plausibly compared with those derived from due application of the **IBO** criteria and procedure.

[375] As mentioned previously, appropriate, norm-referenced criteria applicable to the specific context of language production within the **IBO** programme researched, are not available, requiring special devising were they to be used as a supplementary comparator for triangulating the research.

See *Note No. 358.*

[376] As will subsequently be seen, partial resolution of this problem was proposed through the use of a further refinement of the model, allowing supplementary triangulation for improved validity, and better purchase on the problems of reliability.

[377] Given the bounds set for the project, it has not been possible to research for example, the processes by which assessor judgements are formed, both globally and within each discrete assessment category. Such would be desirable for any extension of the project in future research.

However, the research still serves as an exploration of material gathered as empirical evidence, and processed in order to produce data capable of illuminating the problems of understanding and measuring features of authenticity, as identified. The knowledge derived in this way, may indeed be predominantly heuristic and individualistic in status, although for that reason, it is suggested, not without use in investigating the key issues involved.

[378] It should be noted in this respect, that certain anomalous cases, occurring in the sampling of evidence, remain 'aberrant' despite experimental assessment. The design and procedures of the model are thus challenged in validity, an aspect that requires further, separate description, analysis and discussion, provided in Chapter 6.

Assessing Reading and Writing

[379] This sample varied for each session, between a maximum of 154 and minimum of 150.

The productions were by candidates registered for *Group 2 Languages: French Language B, Standard Level* and allocated to the *Assistant Examiner* by the **IBO**. This was as envisioned in the original project proposal, following pilot research, completed earlier in order to test for feasibility.

[380] The moderated samples number 20 annually, for each of the discrete, May examining sessions from 2000 to 2002.

[381] See **IBO** (2001a; 2001d), *op. cit.*

[382] See Chapter 2, p. 55, *et sq.*

[383] It is not doubted that further, future control of other, recognised variables would improve the generalisability of research conclusions.

[384] This moderation of assessor judgements by the supervising *Examiner* and **IBCA** employee is made in the interests of measuring inter-rater reliability and determining a mathematical co-efficient for each *Moderator* and *Assistant Examiner.* The result is applied in any eventual adjustment of final scores as compensation for irregularity. In turn, the work of such *Team Leaders* is further moderated on the basis of a sample of approximately 50 assessed copies of candidate work, and a large sample of copies of moderated work provided by the members of the team. This step in moderation, preceding the work of *Grade Award Meetings*, is completed by the *Principal Examiner.* The resulting co-efficients determined by correlating all the scores obtained, following analysis by linear regression, are compounded to obtain a final correlation factor, *r*, for each examiner, by which their scores are adjusted.

[385] This honoured the ethical commitment made under the research proposal: that is, to dispatch regular reports on progress to the organisation in return for the granting of access to its archives.

The professional colleagues consulted are indicated by name in the *Acknowledgements* that preface this thesis.

Assessing Listening and Speaking

[386] That is, samples of candidate work, selected by the examining centres themselves and involving on average, a ten to fifteen minute presentation and discussion both of a specific topic, personally chosen by the candidate from one of the three general theme areas of the programme of the '*Exploration of Change*', '*Exploration of Groups*', and '*Exploration of the World of Leisure*', and consequent, more general conversation with the teacher-internal assessor.

[387] However, for data produced in 2002 and 2003, reference was made to the refined version of the assessment grid, as shown in **Appendix 3.**

See Van Lier (1996), *op. cit.*

[388] This makes use of Microsoft's *Excel* Spreadsheet computer programming,

[389] In this context, 'short-term, longitudinal stability' is taken to indicate the period of time required to complete the assessment and evaluation of one batch of candidate production, as identified by **IBCA** and allocated to the *Moderator* and *Examiner* concerned. Typically this would involve a period of between three and four weeks' part-time devotion to the exercise. 'Mid-term longitudinal stability' refers to the stability of interpretations and judgements over a single examining session, measured from the beginning of assessment and evaluation with the receipt of samples from examination centres by the *Moderator* and *Examiner* concerned. It ends with the contextualised description, analysis and discussion of the final stage in the process, established through publishing the *Subject Report* for the language, level and evaluation session.

Typically this period lasts from between six and seven months in any given year. 'Long-term longitudinal stability' is used to refer to the stability of interpretations and judgements made by the researcher in the interests of creating perspective and means for measuring and evaluating such stability, through constant reconsideration, re-assessment and re-evaluation of the samples of language production received and retained throughout the course of the research. That is, the period of time represented stretches formally over a period of three years from the commencement of research, and informally longer, over the period of the researcher's involvement with such material in the context of professional, **IBO** employment.

As is explained, alternative measures for such 'stability' are retained by **IBCA** in archival material, taken for organisational purposes to establish the validity and reliability of the *Moderator* and *Examiner* in question. The data is employed in the determination of moderation factors, or correlation co-efficients, identifying degrees of stability, or 'consistency' of interpretation and judgement, as well as tendencies to either 'generosity' or 'severity' in the allocation of scores to individual samples of language production.

[390] Not least for validity, it requires detailed controls for inter-rater reliability as measurements of additional variables.

See the discussion of the problems raised by this dimension in, for example, McNamara (2000), *op. cit.*

[391] This was based on the evidence of moderation of samples of the researcher's assessment, made in fulfilment of **IBO** duties for May 2000, May 2001 and May 2002.

In the case of *Internal Assessment* and with application of a linear regression, statistical manipulation, acceptable moderation factors of $r = 0,99$ (May 2001), and $r = 0,94$ (May 2002) were derived, where $r = 1$ is a perfect figure. For *Written Production* similarly acceptable moderation factors of $r = 0,97$ (May 2000), $r = 0,97$ (May 2001), and $r = 0,98$ (May 2002) were derived.

It should be noted that in each exercise, different **IBCA** *Moderators* and *Team Leaders* were used, with different moderation factors applied to their own assessments and assessments of the samples provided by *Internal Assessment Moderators* and *Assistant Examiners.*

(Communication to the researcher in conversation with the **IBCA** *Director of Assessment*, in August 2002.)

The scores from which the data are derived are represented graphically and recorded in **Appendix 5**.

[392] The actual figures resulted in 50 examples being considered in 2001, 54 in 2002, and a further 54 examples analysed in 2003, without producing significant new evidence requiring further description, analysis and discussion.

[393] See **IBO** (2001a; 2001d), *op. cit.*

Further Developments of Method

[394] The recommended procedure of the **IBO** is recounted in Chapter 2 and thus not reiterated here.
See Chapter 2, p. 58, *et sq.*

[395] On completion, the remaining results were also dispatched directly to **IBCA** in the required form, for further processing according to the procedures of *Moderation* described in Chapter 2.
See Chapter 2, p. 59, *et sq.*

[396] Most usually, these cases concerned samples where the criteria could not be applied for reasons of deficiency in the recording of oral production, as noted previously.
See **IBO** (2001a,b,c), *op. cit.*

Observation and Recording of *Grade Award Meetings*

[397] For data-collection, devising a draft schedule for general observation of these meetings was initially proposed. As had been determined *a priori* at the outset of the project, this was to be used in pre-structuring, and thence selectively focussing observer attention on aspects of each meeting, likely to relate to, and prove informative for the objectives of the research. The schedule devised considered aspects of authenticity as contextualised, communicative interaction within the constraints imposed by task-design, standardisation or task-equivalence procedures and their outcomes, language-production, assessment, moderation and evaluation.

Originally, it had been proposed to record the proceedings of the first meeting on audio-tape in their totality, for later transcription and analysis, allowing greater scope in identifying items of research relevance. In this way, it was anticipated that limitations, omissions and possible distortions attributable to the predetermination of observer perspective through the chosen strategy of data identification, selection, collection and recording, could be balanced and re-analysed on later occasions, with reference to fixed recordings. Reports drawn from this data-base could subsequently be shared with participants in the process for further comment, and modification, given any omission, misinterpretation or misrepresentation. Such a procedure would serve as a counterweight, balancing undesirable effects created by closures of data-identification and collection through the structuring of the recording methods adopted. The compression of data into a framework of concepts determined solely by the researcher, would thereby be re-adjusted. The outcomes of this approach to collecting empirical evidence through observation were finally to be discussed and further recorded on audio-tape, in semi-structured interview of the participants at the meeting. The structure for interviews was to be determined as rapidly as feasible, after the observation of each stage of the meeting, and with a focus highlighting the observer's understanding of points of interest that merited comment, as they arose during the proceedings. For ethical reasons, the whole process was to be subject to the prior approval of the **IBO** and of the **IBCA** personnel concerned.

In the event of the meeting for *French, Language B*, and given the previously unexpected absence of the *Subject Area Manager* for *Group 2 Languages*, the *Chief Examiners* in attendance expressed the desire *not* to be recorded on audio-tape in their discussions. This was due to the extra stress under which they were working, with less specialist and experienced guidance and advice available from the **IBCA** managers present, and a pressing need to respect tight deadlines for the completion of the moderation and evaluation procedure. It was clear, and indeed made a condition of access to the meeting for research purposes that the observation, recording and interview processes of the project should in no way intrude on this unusual set of circumstances. However, such intrusiveness as an *observer,* rather than as a researcher, was noted in comment to the researcher from the *Chief Examiner.* This was exacerbated on account of the absence of the *Subject Area Manager*, through illness, and entailed greater scrutiny of all processes, with less taken for granted, than normal.

As a preliminary exploration of **IBCA** moderation and evaluation procedures, from which descriptions would be derived, permitting future analysis and renewed observation according to more precisely-defined criteria, at later dates, it was decided on reconsideration of the objectives of observation, that the more rigorous strategy originally proposed was prematurely prescriptive and unduly limiting.

[398] In the case of *French,* this was as teacher and fellow-examiner.

[399] See **IBO** (2000c), *op. cit.*

[400] Given the sensitive confidentiality of the process and of the content of meetings, the *Chief Examiners* present were unwilling to be tape-recorded in interview, as originally proposed. It was also felt that such recording would serve to increase stress at meetings that were in themselves, highly-stressful, as examiners sought to achieve consensus in judgements within highly restricted periods of time.

See *Note No. 397.*

[401] Private communication to the researcher from the *Chief Examiner.*

[402] The findings of the exercise are described and discussed in Chapters 2 and 6. They are based on the official reports drawn up within two weeks of the meetings, and dispatched to **IBCA** in accordance with required procedure.

[403] In the eventuality and in both cases, no requests were received from any present for cross-checking the notes taken. The separate meetings held at the same times with *Director of Assessment* ensured that no material of confidential import for **IBCA** had been recorded.

[404] Indeed, given the researcher's limited competence in German, it is possible that greater attention, focusing on the precise meanings of vocabulary used at this meeting, ensued. The researcher's fluency in French may be partially suspected as a source of potential unreliability. In effect, a tendency to presume shared meanings amongst those attending the earlier meeting at **IBCA** in December 2000 may thereby have been favoured.

Research Data and the Design and Standardisation of Tasks

[405] In this, *French Ab Initio, Language B* and *Language A2* come within the framework of the *Group 2 Languages* programme. The problems of categorisation are indicated in the material research and are of interest, even though for the greater part, they concern the design and standardisation of *Paper 1: Text-Handling.*

[406] A description of relevant data, as identified in the documentation consulted at **IBCA** together with discussion is contained in Chapter 2 and further discussed in Chapter 6.

[407] The curricular description of this programme component is contained in the **IBO** publication: *The Diploma Programme: Group 2 Languages* (in English and French versions), (1996b). It may be noted that it in the main, this description concerns the detailed administrative guidelines required by candidates for producing and recording internal assessment material, in conjunction with their teacher-assessors. Sections pertaining to conceptualisations of authentic language production and to the procedures influencing such linguistic production, its recording, assessment and moderation have been described and discussed in Chapter 2.

See **IBO** (1996b), *op. cit.*, pp. 28 – 3.

The Longitudinal Dimension of Data-Collection and Processing

[408] With written production, experimental assessments were completed at differing times as noted.

CHAPTER SIX: Evidence

Preface

[409] See Chapter 1, pp. 25 – 28.

IBO Publications and Documentation for Internal Use in Formal Assessments

[410] The necessity for strictest confidentiality with material referring to the 2000 and 2001 examining sessions has now diminished of course, given that the administration of the examination and its assessment, moderation and evaluation is now complete, with the results published in the public domain and the time-limits for queries and appeals now lapsed.

[411] See Chapter 2, p. 45, *et sq.*

[412] See Chapter 2, p. 47, *et sq.*

[413] The role of the latter is defined explicitly as:

> "essential [.....] in ensuring the academic integrity of **IB** assessment within each subject. [*Subject Area Managers*] are involved throughout the process of examination paper production, providing guidance to examiners and other members of the team to ensure that the question papers take account of the nature of the **IB** candidature and are a fair and appropriate reflection of the **IB** programmes which they aim to assess."
> See **IBO** (1996b), *op. cit.,* p. 4.

Task-Design and the Editing of Authentic Texts as Resources

[414] For research purposes, this is understood as material produced for native-speaker readers and audiences in a context concerning neither explicit processes of acquiring language for its own sake, nor the assessment of such language acquisition.

[415] Exceptions are foreseen and categorised as occurring through possible breaches of internal organisational security, or as effects in either irregular relationships between any of the group of examination-designers, school administrators, teachers and the candidates of the examinations, or irregular procedure in observing the rubrics and parameters of assessment.

As reported and discussed in Chapter 2, candidates themselves largely determine the content of *Internal Assessment* through appropriate, personally-chosen, oral presentations, debate and interview. Hence, this is omitted from present discussion.

[416] Only the instructions relating to *Paper 2 Written Production* are directly relevant to the project, and duly reported here. As revealed, the form and content of *Paper 1 Text-Handling* are pertinent insofar as general themes presented may relate to tasks posed in *Paper 2*. From scrutiny of documentation relating to examination design, questions of authenticity in *Paper 1* may be understood as concerning in the main, manipulations of linguistic material presented to candidates, ensuring conformity with the **IBO**'s discrete rubrics for this particular examination paper and level.

[417] **IBCA** (2001c), *op. cit.*, p. 4.

[418] It is interesting to note that the term "sophistication" is not qualified. It is therefore ambiguous whether it entails sophistication of task response, form, content or language, though in the absence of qualification, and given the design of *Assessment Criteria*, as presented later in the chapter, all are probably to be taken as comprehensively intended.

See **IBCA** (2001c), *op. cit.*, p. 4.

[419] **IBCA** (2001c), *op. cit.*, p. 4.

[420] **IBCA** (2001c), *op. cit.*, p. 4.

[421] **IBCA** (2001c), *op. cit.*, p. 4.

[422] A list of possible exemplars is included. **IBCA** (2001c), *op. cit.*, p. 5.

[423] **IBCA** (2001c), *op. cit.*, p. 5.

[424] **IBCA** (2001c), *op. cit.*, p. 5.

[425] It is interesting to note that as a tentative example, "letters about holiday plans might be appropriate as a *SL* question but not as a *HL* one". The implications for issues of authenticity will be discussed in later sections of this chapter.
See **IBCA** (2001c), *op. cit.*, p. 5.

[426] **IBCA** (2001c), *op. cit.*, p. 5.

[427] The format of the examination and its tasks are recounted in Chapter 2.
See also **IBCA** (2001c), *op. cit.*, p. 5.

[428] **IBO** italicisation. **IBCA** (2001c), *op. cit.*, p. 5.

[429] **IBCA** (2001c), *op. cit.*, p. 5.

[430] **IBCA** (2001c), *op. cit.*, p. 5.

[431] See **IBCA** (2001c), *op. cit.*, p. 1.

[432] **IBCA** (2001c), *op. cit.*, p. 1.

[433] **IBCA** (2001c), *op. cit.*, p. 1.

[434] See Van Lier (1996), *op. cit.,* Chapter 6.

[435] The arrangements for *Internal Assessment*, where candidates largely determine the form and content of their oral presentations should be contrasted with the rubric-design for *Written Production.* The former evidently and effectively grants unrestricted scope for authentic expression to candidates themselves.

[436] These are examined at a later stage in the present chapter.

[437] Authenticity here may be taken as *Creator Authenticity* and *Authenticity of Interaction*, as defined by Van Lier.
See Chapter 5, pp. 135, *et sq.*
See also, Van Lier (1996), *op. cit,* Chapter 6.

[438] The former are explicitly privileged as the primary source from which form and content for examinations and their assessment criteria are derived.

[439] The following for example, were noted from documentation and analysis reported in Chapter 2 as contextual background:

- the differentiation between *Standard* and *Higher Level* is one of "sophistication", rather than anything else;
- there is an explicitly stated "difference of expectations" between the levels, although these differences and expectations remain implicit;
- the requirement that examination designers and standardisers consider the 'suitability' of tasks according to the specified level underlines differences of expectation that continue to remain implicit.

It is interesting to note that as a tentative example, "letters about holiday plans might be appropriate as a *SL* question but not as a *HL* one". The implications for issues of authenticity will be discussed in later sections of the present report.

See **IBCA** (2001c), *op. cit.*, p. 5.

[440] See also, Chapter 2, p. 46, *et sq.*

[441] **IBCA** (2001c), *op. cit.*, p. 1.
See also, Chapter 2, p. 46, *et sq.*

[442] **IBCA** (2001c), *op. cit.*, p. 1.

[443] **IBCA** (2001c), *op. cit.*, p. 1.

[444] The researcher's italicisation and emphasis.

[445] See also, Chapter 2, p. 49, *et sq.*

[446] See Hymes (1974), *op. cit.*

[447] These are reported in Chapter 2, p. 46, *et sq.*

[448] **IBCA** (2001c), *op. cit.*, p. 3.
See also, Chapter 2, p. 46, *et sq.*

[449] This is described in Chapter 2, p. 54, *et sq.*

[450] For ethical reasons, the security concerns and interests of the **IBO** and in conformity with agreements under which primary data could be selected and collected, the notes from which this section is derived were shown on completion to the **IBCA** *Examination Papers Officer* from the **CEPP** department, and the *Director of Assessment*, with invitations to comment. No alterations were made and permission was granted to make use of their content for research purposes. A photocopy of the full set of notes was produced and passed to the **IBCA** *Subject Area Manager* concerned.

[451] See Chapter 2, p. 54, *et sq.*
See also *Notes Nos. 57 and 61.*

In the case of the French language, where gender is specified in adjectival morphology, a specific problem is posed by the manner in which candidates may be addressed.

[452] The collection of further examples would be desirable for generalising in conclusion. Such has not been possible within the framework of the research timetable. The examples derived from the preparation of the examination for May 2001 are thus restricted to the design of the six tasks offered.

[453] The issue of authenticity for composing personal diary entries *in a foreign language* as the target language of the examination, is evidently central to the concept of *Creator Authenticity* as the expression of 'self' and defined by Van Lier.

See Van Lier (1996), *op. cit.*

[454] It may be noted here that the change offers a possible explanation for a perceived problem in assessing responses to this task. The inclusion of a purpose of "going to university", may be taken firmly to anchor the task as reflection on feelings about an anticipated event in the near future. With the removal of this specific purpose, for reasons of appropriate contextualisation for those not planning to attend university, the ambiguity not only allowed space for freer candidate interpretation of possibilities for response content, but favoured a mis-reading of the task as a reflection on leaving the family in the recent past. Such a mis-reading proved common and incurred a penalty on assessment - a feature further discussed later in this report.

The *Internal Assessment* Component

[455] See Chapter 2, pp. 44 - 45.

[456] See Chapter 1, pp. 25 - 28.

[457] See **IBO** (1996b), *op. cit.,* pp. 28 – 34.

[458] See **IBO** (1996b), *op. cit.,* p. 28.

[459] In this way it may be surmised, the problem of 'familiarisation', promoted by repeated practice of identical activities, may be significant with regard to the criteria set.

[460] Although this requirement appears to run counter to the previous criterion, cited and discussed in *Note No. 42,* it is not an obligatory one. It should be recalled that examining centres may exercise discretion in the best interests of their candidates as to the choice and allocation of an *Internal Assessment* interlocutor.
See Chapter 2, p. 44, *et sq.*

[461] The latter two requirements form necessary constraints on the format and context of the individual oral, allowing moderation and evaluation across a range of candidates, teacher-assessors, examining centres, moderators and evaluators. Given that candidates are the only actors in the process significantly to be affected by its outcomes, and that moderation of the performance of all other actors is a necessary requirement for equity, such restrictions on the ability of candidates to 'create' communication according to their own agenda, (a significant aspect of the 'autonomy' required for authenticity in Van Lier conceptualisations) are not seen as significant for research purposes. Indeed, response to a given context and cultural setting, as chosen by the candidate on registration for the **IBO**'s examinations, is in itself a central constituent component of authenticity.

[462] See **IBO** (1996b), *op. cit.,* p. 28.

[463] See **IBO** (1996b), *op. cit.,* p. 30.

[464] See **IBO** (1996b), *op. cit.,* p. 30.

[465] See **IBO** (1996b), *op. cit.,* pp. 28 and 31.

[466] See **IBO** (1996b), *op. cit.,* p. 28, pp. 31 and 32.

[467] See **IBO** (1996b), *op. cit.,* pp. 28 and 32.

[468] See **IBO** (1996b), *op. cit.,* p. 28.

[469] See Chapter 5, p. 136, *et sq.*

[470] In this case 'balance' may be presumed as indicating linkage between the concerns of 'self' and 'other' in a combination of subjectively personal and culturally-conditioned ways. Hence, it involves relevant activities that 'cannot be externally evaluated'. The understandings of possible meanings for these vague terms require further research into the views of programme designers, assessors, moderators, evaluators, teachers and candidates that has not been undertaken beyond early, pilot investigations, given the practicabilities of scope for the present research.

Criterion Descriptors for *Internal Assessment*

[471] Summarised from the *Guide to the Programme for French Language B, Standard Level,* **IBO** (1996b), *op. cit,.* pp. 49 – 51.

The criteria establish intermediate descriptors between these extremes, as will be seen later in this chapter.

[472] See Chapter 6, pp. 173 - 177.

[473] See subsidiary research questions, Chapter 1, p. 28.

Paper 2: Written Production

[474] Descriptions of this component are provided in Chapter 2.
See pp. 44, *et sq.*

[475] As will be seen and was partly explained in Chapter 2, failure accurately to read the instructions for any given task may lead to substantial penalty in response, notably under the criteria for *Task* and *Message*, as well as those for *Presentation.*

[476] See **IBO** (1996b), *op. cit.,* pp. 17 *et sq.*, for examples of the types of text possible for setting in examination.

[477] See **IBO** (1996b), *op. cit.,* pp. 6 – 23.

[478] **IBO** (1996b), *op. cit.,* p. 6.

[479] **IBO** (1996b), *op. cit.,* pp. 9 – 10.

[480] **IBO** (1996b), *op. cit.,* p. 6.
See Chapter 2, p, 40, *et sq.*

[481] **IBO** (1996b), *op. cit.,* p. 7.

[482] See Chapters 2 and 4, p. 45 and pp. 117 – 118.

[483] See **IBO** (1996b), *op. cit.*, p. 8.

These aims are further explained in the specific detail of programme *Objectives*, common to both *Standard* and *Higher Levels*, as described in Chapter 2.

See Chapter 2, p. 39, *et sq.*

[484] See **IBO** (1996b), *op. cit.*, p. 16.

[485] See Chapter 5, p. 136, *et sq.*

[486] The effects of such problems are discussed later in the present chapter.

[487] As in *Internal Assessment*, 'balance' may be presumed to indicate linkage between the concerns of 'self' and 'other' in a combination of subjectively personal and culturally-conditioned ways. Hence, it involves relevant activities that "cannot be externally evaluated". Understanding possible meaning for these vague terms through further research into the views of programme-designers, assessors, moderators, evaluators, teachers and candidates was not continued beyond pilot research, for reasons of practicability.

Criterion Descriptors for *Paper 2: Written Production*

[488] See also Chapter 2, p. 36, *et sq.*

[489] See Chapter 1, pp. 25 - 28

See also the key questions for research, defined on p. 27.

[490] These are of course, matched with *General Grade Descriptors* and referenced to the **IBO**'s seven-point scale, as described in Chapter 2.

See Chapter 2, p. 33, *et sq.*

[491] This data is mainly derived from the *Subject Guides: Language B* in the French version, since this is the one to which assessors for *French Language B* are expected by **IBCA** to refer.

IBO (1996b), *op. cit.*, pp. 38 – 44.

Problems of translation from English to French, or **vice versa** do not appear relevant in this instance, although cross-referenced research of the two respective publications for *Subject Guides: Language B* has been completed. A composite summary is given in the table presented, with information derived from the *Subject Guide: Language B* in French, predominating.

[492] In this sense, administrative decisions for registering candidates at one or the other level may thus determine gradations of success or failure, as noted in Chapter 2.

See pp. 36 – 38.

[493] It should be noted that according to the procedure for assessors, the choice between any two values in a given category is left to the professional judgement of the individual assessor. However, advice given in conversation between examiners and at **IBO** training sessions for teachers, suggests that relative severity in judging one criterion may be compensated with relative generosity in another.

Researcher's understanding following attendance at **IBO** organised teacher-training workshops.

[494] In cases of doubt on the part of the assessor, the procedure designed by the **IBO** requires consultation of the more detailed descriptors accompanying each category. Hence for example, in the case cited previously, detailed descriptors add the refinement of understanding that "on the bare limits of adequacy" should be interpreted as meaning that first, the ideas presented are "**generally** superficial", and secondly, are either "**repetitive**" or irrelevant to the task proposed. Whereas "superficially completed" (or in the official English version: "never go[ing] beyond the obvious") is to be interpreted as meaning that first, the ideas "**can sometimes be** superficial", and secondly, are either "**sometimes repetitive**" or "**sometimes irrelevant**" to the task proposed. It should be noted once again that these data are derived in the main, from the French language edition of the *Assessment Criteria* (in accordance with the expectations of the **IBO** for the assessment of texts produced for *French Language B*), where ambiguity may be created by the use of the intensifier "à peine adéquate", conventionally translated into English as "barely (or scarcely) adequate". The official English language version of the same criteria gives "barely adequately carried out".

See **IBO** (1996b), *op. cit.*, p. 39.

[495] In this context, the general criteria for the minimum aggregation of grades in awarding **IB** *Diplomas* should be held in mind. With Grade 4 gained in any individual component, no additional requirements need 'normally' be satisfied, beyond the scoring of a minimum of 24 points from a maximum of 45. However, the list of "failing conditions" for the award of a final *Diploma* by the **IBO** is detailed, complex and extensive, with as many as 39 discrete categories defined. These 'failing conditions' are listed in the published *Vade Mecum* for the *Diploma Programme*.

See **IBO** (1996a), *op. cit.*, Section G 16.9, pp. G 32 – G 33.

[496] Once again problems may be seen to arise from the ambiguity created in translation of the French expression "à peine".

See *Note No. 494.*

[497] See **IBO** (1996b), *op. cit.*, p. 39.

[498] 'Ambiguity' in this context indicates greater freedom for individual assessors to decide idiosyncratic meanings, thus extending conventions and introducing a greater element of subjectivity in the assigning of point-value categories to a given text production.

[499] It should be noted as before, that according to **IBO** procedure, the choice between any two values in a given category is left to the professional judgement of the individual assessor.

[500] In summarising the French versions of the general criteria in English, interpolations are made by the researcher in order to make implicit assumptions of meaning explicit, thus better illustrating gradations in intensity and conceptualisation.

[501] Once again, according to the procedure for assessors, the choice between any two values in a given category is left to the professional judgement of the individual assessor.

[502] This is according to *Chief Examiners' Subject Reports.*

See **IBO** (2000b, 2001f), *op. cit.*

[503] See **IBO** (1996b), *op. cit.*, pp. 41 and 44.
It may be noted for example, that the *Chief Examiners* in their *Subject Report* for *French, Language B, Standard Level, Paper 2: Written Production* for the November 2000 session of the examination make the general comment that:

> "It would be good policy to insist that candidates write out their work appropriately, without crossings out and in a clearly legible manner."
>
> (Researcher's translation from the original French, reading as follows:
>
> "Il serait bon d'insister auprès des candidats pour qu'ils écrivent leur travail proprement, sans ratures et de manière bien lisible".)
> **IBCA** (2000b), *op. cit.*, p. 7.

However, in the case of work struck out in examples of response to the tasks posed in *Paper 1: Text-Handling,* it was noted at the December 2000, *Grade Award* meeting that such work could be considered in clarifying assessments of cases close to a grade boundary.

[504] In this context it may be noted that whereas languages such as English and French may treat punctuation more as an aspect of presentation, with bearing on nuances of meaning, languages such as German consider this a function of grammar where precision is required for correctness. In all cases, punctuation is an element that forms part of the likely expectations of an 'educated' reader in any informal, impression-based assessment of a written text.

[505] See pp. 184 and 188.

[506] See **IBO** (1996b), *op. cit*, p. 38.

[507] Their content has been translated into English purely for research purposes, since English Language versions are not used in this domain, and hence are not generally available. However, it is assumed that significant deviation between versions is unlikely. Indeed in this context, such would be irrelevant.

[508] The methods adopted for the preparation and further use of this aspect of the research are described and discussed in the previous chapter.
See Chapter 5, p. 158, *et sq.*
The contents of the Reports are summarised in Chapter 2.
See pp. 33 - 36.

Supplementary Documentation

[509] This was however dated February 2002.

[510] See **IBO** (2001a), *op. cit.*

[511] **IBCA** (February 2001; February 2002), internal circulars drawn up in French for all *Assistant Examiners* for the relevant programme.

[512] Communication to the researcher from **IBCA**, composed by the *Chief Examiners.*

[513] See Chapter 6, p. 191.

[514] See Gipps (1994), *op. cit.,* p. 105.

[515] See Chapter 6, p. 197.

[516] The conversions are made in accordance with data published in the appropriate *Subject Reports.*

In these reports the following scales are given, with initial digits indicating the range of points scored. These are equated in value to points on the **IBO**'s seven-point scale.

Thus for May 2000 at *Standard Level*, a score of 0 to 4 points is equivalent to Grade 1 in attainment. 5 to 9 points = Grade 2; 10 to 13 points = Grade 3; 14 to 17 points = Grade 4; 18 to 22 points = Grade 5; 23 to 26 points = Grade 6 and 27 to 30 points = Grade 7.

For May 2001, they are respectively: 1 to 3 points = Grade 1; 4 to 7 points = Grade 2; 8 to 13 points = Grade 3; 14 to 17 points = Grade 4; 18 to 23 points = Grade 5; 24 to 26 points = Grade 6 and 27 to 30 points = Grade 7.

These particular evaluations are determined at the appropriate *Grade Award Meetings*, the procedures of which have been identified, described and discussed.

However, the rationale for alterations in grade values made between sessions with identical tasks and rubrics, thereby assumedly invariable at a given level, has not been extensively researched. From sponsored training sessions, it could be inferred that the desire to prevent stable and accurate assumptions, establishing a place for norm-referencing, could be an **IBO** concern in devising and applying such policy.

The documents consulted exist in French versions only. References used in the body of the report have been translated or summarised in English by the present researcher.

[517] The overall grade-boundaries are published for each formal assessment session in the appropriate *Subject Report*, as previously described.

[518] It should be recalled in this context that the overall grading for this component accounts for 30% by value of the total for the subject and level, as stated in the relevant *Subject Guide.*

See **IBCA** (1996b), *op. cit.*, pp. 25 – 26.

[519] Researcher's translation from the original French.

[520] See *Note No. 516.*

[521] See **IBO** (1996a), *op. cit.*

[522] This is quoted from **IBCA** sources in French, and cross-referenced as equivalent, without significant variation, to the official English language version.

[523] These are discussed later in the present chapter.

The *Subject Report* for *French Language B*: November 2000

[524] See **IBCA** (2000b), *op. cit.*

This exists in a French version only. References used in the body of the report have been translated or summarised in English by the present researcher.

[525] In this way for example, reporting on aspects of the examination for *Paper 1: Text-Handling*, is excluded from the present description and discussion inasmuch as its details do not relate to questions of authenticity concerning other *Diploma Programme* components.

[526] See research aims and questions, Chapter 1, pp. 25, *et sq.*

[527] It should be noted that all *Assistant Examiners* are required as a part of their official duties, to report to the *Chief Examiners* according to an **IBCA**-determined agenda for compiling *Subject Reports.* Teachers are invited to do so by completing and returning a questionnaire that accompanies each examination for this purpose.

[528] See **IBCA** (1996b), *op. cit.*, pp. 3 and 4.

[529] The task proposed in this case was to compose a *guide* for a readership of peers, evaluating the experience of studying within a *Diploma Programme* curriculum.

[530] An example given relates to the use of performance enhancing drugs in professional sport, related to the content of *Text B* in *Paper 1* and *Task 1* of *Paper 2.*

[531] As in *Tasks 3* and *4*, where the respective requirements are to compose the texts of speeches for oral delivery to the target audience identified, or a 'typical' editorial for publication in a newspaper. The *Chief Examiners* report that in the former case, conventional essay form is 'inappropriate', since task specific forms of opening and closing a speech are expected. Convincingness is, in part, created in the mind of the reader, or audience, through the 'successful' adoption of a tone of enthusiasm. In the latter case, the form and tone of an editorial is explicitly categorised as the statement of personal points of view, vividly presented in a personal style, with striking exemplification, development in a single direction and the elimination of discordant 'voices'. An editorial is seen as a statement of prior conviction aimed at persuading the reader to adopt the point of view presented.

See **IBCA** (2000b), *op. cit.*, p. 4.

[532] These refer to the concepts of *Creator Authenticity* and the need for choice in the construction and communication of 'self', *Authenticity of Purpose, Authenticity of Interaction*, and so forth, as identified and classified by Van Lier (1996), *op. cit.*

[533] The examiners state: "le rythme devient effréné et adieu l'authenticité et l'aisance, cela ressemble aux *Exercices de Style* de Raymond Queneau."

See Queneau (1947), *op. cit.*

The latter is a reference to a celebrated text of this French author, whereby a trivial story is retold in 99 differing 'styles' of language and textual format.

See **IBCA** (2000b), *op. cit.*, p. 5.

[534] 20 responses are summarised as follows: 3 centres found the examination "easier"; 13 found it "similar"; and 4 found it "slightly more difficult".

See **IBCA** (2000b), *op. cit.*, p. 5.

[535] The total number distributed has not been determined, though the present sample, from communication to the researcher by Anne Scott, the *Chief Examiner* at the time of the December 2000 *Grade Award* meeting, was deemed "fairly small".

[536] Translation by the researcher of the French language statement that: "le niveau de difficulté de l'épreuve 2 était approprié aux connaissances des candidats". It should be noted that the concept of "connaissances" has been translated as "knowledge and experience" in this context.

See **IBCA** (2000b), *op. cit.*, p. 7.

[537] See **IBCA** (2000b), *op. cit.*, p. 7.

[538] This was noted in reporting observation of the relevant *Grade Award Meeting* of December 2000 and is contrary to the rubrics of the examination.

See Chapter 2, p. 44, *et sq.*

[539] To summarise, *Task 4* required the composition of an editorial and *Task 6*, a project proposal. *Text C* of *Paper 1* concerned difficulties in friendship relations and *Text A* concerned the topic of tobacco smoking.

[540] Characteristic examples are given from assessing the first task proposed, where candidates were required to write a letter to a friend imagined as importuning the author for help with schoolwork, in order to say "no" to future requests for help. The *Chief Examiners* comment that candidates who related the task to a personal experience of preparing for the examination at hand and who contextualised the text production with explanations of why they were writing, within the imagined scenario, were often very successful in producing 'convincing' communications. In this way, the requirements for 'authenticity' as elaborated by Van Lier (1996) for example, appear happily fulfilled.

See Van Lier (1996), *op. cit.*

However, others were judged as less successful in producing texts that lacked sufficient, explicit explanation of the author's varying motivations and points of view. Such responses were deemed wanting either in "validity", or "convincingness", or were at times psychologically inconsistent and contradictory. Many 'weak' responses were also inappropriately expressed in cultural terms, being framed in formal language, when an informal style and choice of vocabulary would have been expected. Many examples are quoted by the *Chief Examiners* as evidence in support of such claims.

See **IBCA** (2000b), *op. cit.*, p. 8.

[541] In this context, an example is given of candidates reproducing material remembered from *Paper 1*, where the dangers of smoking tobacco are listed without adaptation to the requirement of the task to devise a brochure ***advising*** readers on how to give up the habit.

See **IBCA** (2000b), *op. cit.*, p. 8.

[542] For example, *Task 2* on the dangers of tobacco smoking is cited as requiring a certain amount of technical vocabulary that readers might expect to be used.

See **IBCA** (2000b), *op. cit.*, p. 8.

[543] The example is given in the case of the third task proposed, relating the need to respect expected social and cultural forms of writing letters to anonymous and unknown recipients in formal contexts.
See **IBCA** (2000b), *op. cit.*, p. 9.

[544] See under *Criterion A: Task and Message.*
See **IBO** (1996b), *op. cit.*, pp. 39 and 42.

[545] In this context, 'authenticity' appears to mean 'appropriateness' of language and register, integrated with the ideas presented. It is presented as an ideal, the attainment of which should serve as the goal of successful examination text-production and is worth quoting. (The researcher's translation follows.)

> "[.....] les enseignants doivent insister [.....] sur la rédaction d'un plan de la tâche contenant la liste des idées principales et des éléments nécessaires au sujet at au destinataire (par exemple le registre, les termes de cohérence lexicaux appropriés au niveau de langue du candidat). Les candidats auront ainsi toutes les chances d'écrire un texte aussi authentique que possible sans répétitions ni longueurs."
>
> [Teachers should insist on the drafting of a plan of the task, containing a list of the key ideas and of items required for the subject and its public (for example, the register, cohesive devices in lexis, appropriate to the linguistic level of the candidate). Candidates will thus have every chance of writing a text that is as authentic as possible, without repetitions or wordiness.]
> (**IBCA** (2000b), *op. cit.*, p. 10.

[546] See **IBCA** (2000b), *op. cit.*, p.10.

[547] See **IBCA** (2000b), *op. cit.*, p.10.

[548] See **IBCA** (2000b), *op. cit.*, p.10.

[549] See the section entitled *Généralités*, **IBCA** (2000b), *op. cit.*, p. 10.

[550] This relates to criteria identified for *Finder* and *User Authenticity.*
See Van Lier, (1996), *op. cit.*

[551] This relates to criteria identified for *Creator* and *User Authenticity.*
See Van Lier, (1996), *op. cit.*

[552] This relates to criteria identified for *Creator* and *User Authenticity*, as well as *Authenticity of Context, Purpose* and *Interaction.*
See Van Lier, (1996), *op. cit.*

[553] This relates to criteria identified for *Creator* and *User Authenticity*, as well as *Authenticity of Context, Purpose* and *Interaction.*
See Van Lier, (1996), *op. cit.*

[554] This relates to criteria identified for *Authenticity of Interaction.*
See Van Lier, (1996), *op. cit.*

[555] The entire set of qualities identified may be further subsumed under the criteria for *Existential, Intrinsic* and *Autotelic Authenticity.*
IBCA (2000b), *op. cit.*, p. 11.

[556] Researcher's translation.
IBCA (2000b), *op. cit.*, p. 11.

[557] In this context, the Examiners stress in their *Subject Report* :

> "This aspect of the oral must be underlined as important since it allows measurement of the candidate's capacity directly to mobilise knowledge, experience and linguistic competence, both at *Standard*, and more significantly at *Higher Levels.*"
> (Researcher's translation from the following original:
>
> "Il faut insister davantage sur l'importance de cette partie de l'oral [Section B] qui permet, au niveau moyen at plus encore au niveau supérieur, de mesurer la capacité de l'élève à mobiliser directement son savoir et ses compétences linguistiques.")
> **IBCA** (200b), *op. cit.*, p. 11.

[558] See **IBCA** (2000b), *op. cit.*, p. 11.
The impact of these reiterations, underlining administration requirements for *Internal Assessment*, as published in the relevant *Subject Guide*, is discussed in the following section of the present chapter.

[559] **IBCA** (2000b), *op. cit.*, p. 11.

[560] See Van Lier (1996), *op. cit.*

Oral Language Production for the May 2001/2002 Examination Sessions

[561] See the relevant sections of the document: *The Diploma Programme: Group 2 Languages,* **IBO** (1996b), *op. cit.*, pp. 25 – 34.

Certain problematic cases that did not follow this rubric in its entirety are identified, described in outline, and discussed later in the present chapter.

[562] See **IBO** (1996b), *op. cit.,* p. 12

[563] The format is described in Chapter 2 but may usefully be summarised in this note. See Chapter 2, p. 44.

In interview with a teacher-interlocutor and *Internal Assessor*, a presentation is discussed in some depth, through formal, or informal, exploratory questioning, devised by the self-same interlocutor and *Assessor*. Ensuing conversation is more general and less predictable, though pertinent, personal and conceivably partially-prepared. The whole should last approximately ten minutes, with six to seven minutes devoted to presentation and related discussion, and three to four minutes to general conversation.

[564] See **IBO** (1996b), *op. cit.,* pp. 31 – 32.

[565] See **IBO** (1996b), *op. cit.*, p. 45.

[566] The topic data-base has been approximately and interpretatively categorised by the researcher, and listed in **Appendix 4**.

From observation of **IBO**-organised, official teacher training workshops for familiarisation with the curriculum and promotion of common understandings and standards in order to reduce inter-rater variation in the interpretation and use of the assessment criteria, these general topics are indeed classified by **IBCA** personnel as so broad as to allow virtually any topic to be chosen by a candidate without risk of penalisation for irrelevance.

[567] For future research, such an observation points to the desirability of investigating understandings of the minimum criteria required for assessment-gradings at different levels within the programme (that is, for a scoring greater than zero at either *Standard* or *Higher Level*, and for an accompanying analysis of programme differentiations of requirements at the various scoring levels available in the range, both at *Standard* and *Higher Levels.*)

[568] The design of the criteria, as described previously, almost prevent such from being the case without recourse to mental representations of the minimal requirements on the part of *Examiners,* assumedly derived from prior training and experience.

[569] The examinable component of the programme accounts for 70% of the final 'grading' from which a score on the **IBO**'s seven-point reporting scale (0 representing minimum attainment and 7 representing maximum) is derived. The remaining 30% is devoted to *Internal Assessment*, concerning the listening and oral components of the *Group 2 Language B* curriculum, common to all modern languages.

[570] See for example **IBO** (2001f), *op. cit.*

[571] In the session for May 2001 for example, a particular problem was raised in this respect by the work of a Mauritian student, evidently familiar with Mauritian French Creole, probably competent at an *A2* level and yet entered for the *Language B* examination. The whole issue of native-speakers of a language presenting themselves for assessment within a programme designed to evaluate *foreign* language learning provides more evident examples of the ethical dilemmas created as issues informing teaching and learning, which, as Bachman and Palmer have claimed, form the rationale for all assessment.

See Bachman and Palmer (1996), *op. cit.*

[572] See for example, Hawkins, (1988), *op. cit.*

[573] See Chapter 6, p. 212 - 216.

[574] See Van Lier (1996), *op .cit.*, Chapter 6.

[575] Transcriptions of these interviews have not been included in the body of this thesis, since being in French, they would require translation into English, an act that creates further problems for method that seeks to avoid the inevitable distortion of data occurring in such renderings. The results of assessment and evaluation have been summarised as presently reported.

576 See **IBO** (1996b), *op. cit.*

See also, Chapter 2, p. 44, *et sq.*, where these rubrics are described in detail.

577 Information supplied to the researcher in August 2002 by the **IBCA** *Director of Assessment*.

578 Certain cases lacked correct communication of marks awarded by the *Internal Assessors* concerned and appear scored as zero in the graphs reproduced.

579 The same data has been analysed and recorded numerically in *Table 6.11*, shown in **Appendix 5**.

Written Language Production for the May 2000/2001/2002 Examination Sessions

580 Namely, the sample produced from the May 2001 session of the examination.

581 In this, it should be remembered that the earlier exercise relating comparative assessments of the May 2000 session samples of candidate work, according to the **IBO**'s criteria as well as those derived from Van Lier, was conducted simultaneously.

PART IV

CHAPTER SEVEN: The IBO Programme and Authentic Language Use in Examinations

Preliminary Conclusions

582 In applying equal quantitative weightings for reception and production, explicit rationales explaining and justifying attributions of positivistically-measured value are frequently wanting.

A well-known example of this approach is the English and Welsh GCSE system that assesses listening, speaking, reading and writing discretely. In testing comprehension this is often through recourse to question and answering in the mother tongue, constrained multiple-choice, Cloze-type, sentence-completion exercises (though the latter also feature in the **IBO**'s approach to *Paper 1: Text Handling*) and so forth. Production is often tested through constrained, prompted role-playing for oral assessment and tightly-prescriptive task definition for writing. Both frequently require production of specific items of vocabulary and grammatical structures, with little choice or occasion for the unconstrained expression of self in interaction with listeners and readers, and with norm-referenced criteria applied in the assessment of highly-valued structural, rather than communicative competence. Prescribed, minimum vocabulary lists are published for the purpose of norm-referencing in these situations.

See for example, Oxford Cambridge and RSA Examinations (2003), *op. cit.*, pp. 70 – 71.

Construct Validity in IBO Task-Designs

[583] See *Note No. 582.*

Interpretative Intercommunication

[584] See **IBO** (2000c), *op. cit.*

Positivistic Concerns in Assessment and Authentic Criterion-Referencing

[585] See **IBO** (2000c), *op. cit.*

[586] However, when assessed and evaluated under the relevant scheme, a few cases remain problematic. They challenge construct validity in the design of rubrics, tasks and assessment criteria. They also question procedural reliability, as recounted.

Through comparing sets of common data relating **IBO** outcomes to the results of experimental manipulations, this finding was thrown into relief and is discussed in Chapter 6.

Conclusions and *Internal Assessment*

[587] See **IBO** (1996b), *op. cit.*

[588] See for example, the requirements of typical *GCSE* schemes that are very roughly equivalent by language level to those for *French Language B, Standard Level,* given a similar prior experience of learning by 'typical' students of French as a Foreign Language for four years, or so.

See Oxford Cambridge and RSA Examinations (2003), *op. cit.*, pp. 70 – 71.

See also the definition of the relevant, 'typical' **IBO** *Diploma Programme* candidate in Chapter 2, pp. 39 – 44.

[589] In comparison, other components of the programme, such as *Written Production*, appear more constrained by their distinctive assessment formats. Written productions create largely anonymous relationships between writer and reader, in which the latter are latterly unresponsive.

[590] See **IBO** (1996b), *op. cit.*

[591] See for example, **IBO** (2000a,b, 2001f), *op. cit.*

[592] See Chapter 4, pp. 108, *et sq.*

Conclusions and *Written Production*

[593] See *Note No. 589.*

[594] See Chapter 4, p. 122.

[595] Positivistic, wholly non-interactive schemes such as much of the *GCSE*, often render candidates totally dependent on full accuracy in comprehending target-language response stimuli for the assessment and evaluation of writing. Productions of inflexible, set answers, matching predicted responses supplied in predetermined *Mark Schemes*, without provisions for choice or negotiations of meaning, are typical requirements. In certain cases, rubrics are supplied in a language other than the target in order to stimulate such production.

See for example, Oxford Cambridge and RSA Examinations (2003), *op. cit.*, pp. 70 – 71.

[596] See **IBO** (2000a,b, 2001f), *op. cit.*

Further Unresolved Problems

[597] For illuminating discussion of this concept, see for example, Hawkins (1988), *op. cit.*

Possible Resolutions Indicated by the Research

[598] These terms may be translated as "limited mastery", "average" and "satisfactory", respectively. (Researcher's translation).

The Criteria and Procedures for Awarding Grades

[599] It was observed that through supplementing the variety of perspectives available, these measures were only used to reassure, and clarify cases of ambiguity emerging from criterion-referenced evaluations, as provided by each examiner.

[600] Anecdotal evidence from conversation with various *Chief Examiners* and *Assistant Examiners* would suggest that per *Internal Assessment* or *Written Production* sample, an allocation of twenty minutes for processing candidate work in accordance with the **IBO** criteria and procedures is common. In this context, it should therefore be recalled that the responsibility is for supervising a programme attracting an entry of 1141 candidates at *Higher Level*, and 4325 candidates at *Standard Level*, in May 2001.

[601] In particular, certain styles of questioning for this paper have been shown to be particularly problematic. It has been suggested for future examination designs that they be avoided if possible. Examples reported are: the problem of assessing comprehension through tasks requiring a sequential ordering of information, whereby candidate error occurring early in the sequence may easily have a 'knock-on' effect,

and thereby cast a shadow over construct validity for this method of assessing comprehension; the constant problem with multiple choice questions of differentiating true comprehension from mere guesswork, especially when candidates are encouraged to complete all questions, regardless of the level of their comprehension, and erroneous choices are not 'penalised' through the subtraction of marks.

[602] More particularly, in the case of the November 2000 examination, it was seen that special care needed to be taken in ensuring that the texts chosen for *Paper 1*, and the tasks associated with them, provided sufficient possibilities for differentiation at all grade levels. In this examination, the chance for students at the top end of the range to display sophistication in comprehension, appears to have been limited, thus posing an evident problem for the moderation procedure. The absence of the *Subject Area Manager*, with experience in judging standards across differing years of the examination made itself more evident in this area of the moderation process, thereby lengthening discussion and the time needed to arrive at a valid judgement of grade level boundaries.

Further Prospective Developments for Future Research

[603] See **IBO** (1996b), *op. cit*

The Experimental Research

[604] However, this conclusion may not be practically valid for the early stages of language learning, as represented in the design of the *Ab Initio* programme. Here, it may be asked whether assessment descriptors and categories that avoid reference to structural knowledge and skill are relevant to the needs of basic level authentic communication when a 'foreigner' is included as a partner in interchange.

CHAPTER EIGHT: The Premises of the Research

The Design

[605] See **IBO** (1996b), *op. cit.*, p. 3.

The implicit nature of present **IBO** programmes has indeed been made explicit in revisions for 2002, subsequent to the review process of 1999 – 2002.

See **IBO** (2002 a,b,c), *op. cit.*, p. 3.

[606] For the purposes of data-collection, no apparent limitation of access to relevant sources was encountered during the research process, though either failure comprehensively to identify such evidence, or for the researcher to make appropriate requests for access remains a possibility that partially limits the validity and reliability of the work completed.

It should be recalled that in part, the research design has sought to minimise such constraint, through the copying of regular progress reports to **IBCA** personnel, not only in respect of ethical considerations raised by the research of data remaining the property of others and not available in the public domain, but also for further comment.

[607] See the organisation's website at: www.ibo.org

REFERENCES

ADORNO, T., (1973). *The Jargon of Authenticity*, Northwestern University Press, Evanston, Illinois.

ADORNO, T., (1969), 'Subject and Object', in O'CONNOR, B., (2000), *The Adorno Reader*, Oxford, Blackwell Publishers.

ARGYRIS, C. and SCHON, D., (1974), *Theory in Practice*, San Francisco, Jossey Bass.

BACHMAN, L.F., (1990), *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press.

BACHMAN, L.F. and COHEN, A.D., (1998), *Interfaces between Second Language Acquisition and Language Testing Research,* Cambridge, Cambridge University Press.

BACHMAN, L.F. and PALMER, A.S., (1996), *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford, Oxford University Press.

BLACK, P., (1998), *Testing: Friend or Foe? Theory and practice of assessment and testing,* London, Falmer Press.

BOURDIEU, P., (1977-1984), in J. B. THOMPSON, (ed.), (1991) *Language and Symbolic Power,* Cambridge, Polity Press,

BREDO, E., (1994), 'Reconstructing Educational Psychology', in MURPHY, P., (ed.), (1999), *Learners, Learning and Assessment*, London, Paul Chapman Publishing.

BRITTON, J., (1987), 'Vygotsky's contribution to pedagogical thinking', in MURPHY, P., and MOON, R., (eds.), (1989), *Developments in Learning and Assessment*, London, Hodder and Stoughton.

BRUNER, J., (1986a), 'The Transactional Self', in MURPHY, P., and MOON, R., (eds.), (1987), *Developments in Learning and Assessment*, London, Hodder and Stoughton.

BRUNER, J. (1986b), *Actual Minds, Possible Worlds,* Cambridge, Massachusetts, Harvard University Press.

BRUNER, J., (1999), Audio-taped lecture to the American Educational Research Association, in MURPHY, P., *et al,* (eds.), *E 836: Learning, Curriculum and Assessment: Study Guide*, Milton Keynes, The Open University.

CANALE, M. and SWAIN, M., (1980), 'Theoretical Bases of Communicative Approaches to Second Language teaching and Testing', in *Applied Linguistics,* Vol. 1. No. 1., Oxford, Oxford University Press.

CSIKSZENTMIHALYI, M., (1990), *Flow: the psychology of optimal experience,* New York, Harper & Row.

CHOMSKY, N., (1965), *Aspects of the Theory of Syntax,* Cambridge, Massachusetts, MIT Press.

DODSON, C.J., (1967), *Language Teaching and the Bilingual Method,* London, Pitman Publishing.

EDWARDS, D., and MERCER, N., (1987), *Common Knowledge: the development of understanding in the classroom*, London, Methuen.

FAIRCLOUGH, N., (1989), *Language and Power,* New York, Longman Group UK, Ltd.

FREIRE, P., (1974), 'The Politics of Education', in MURPHY, P., and MOON, R., (eds.), (1989), *Developments in Leaning and Assessment*, London, Hodder and Stoughton.

FREIRE, P. (1996), *Pedagogy of the Oppressed,* London, Penguin Books, Ltd.

GARNHAM, A., (1985), *Psycholinguistics: Central Topics*, London, Routledge.

GIPPS, C.V., (1994), *Beyond Testing: towards a theory of assessment,* London, Falmer Press.

GLASER, R., (1963), 'Instructional Technology and the Measurement of Learning Outcomes: Some Questions', in *American Psychologist,* (1963), Volume 18.

GLASERFELD, E. von, (1987), 'Learning as a Constructive Activity', in MURPHY, P., and MOON, R., (eds.), (1989), *Developments in Learning and Assessment*, London, Hodder and Stoughton.

GOLDFARB, M.E. and ROZYCKI, E.G., (2000), *The Educational Theory of Lev Semenovich Vygotsky (1896 - 1934)*, New Foundations, www.newfoundations.com

HABERMAS, J., (1984), *The Theory of Communicative Action*, Boston, Beacon Press.

HALLIDAY, M.A.K., (1975), 'Language as Social Semiotic', in GRADDOL, D., and BOYD-BARRETT, O., (eds.), (1994), *Media Texts: Authors and Readers*, Clevedon, Multilingual Matters, Ltd.

HASAN, R., (1989), 'The Texture of Text', in GRADDOL, D., and BOYD-BARRETT, O., (eds.), (1994), *Media Texts: Authors and Readers*, Clevedon, Multilingual Matters, Ltd.

HAWKINS, E., (1988), *Modern Languages in the Curriculum*, Cambridge, Cambridge University Press.

HEIDEGGER, M. (1927,1962), *Being and Time,* (translated by MACQUARRIE, J. and ROBINSON, E.), San Francisco, Harper Collins.

HYMES, D., (1971), *On Communicative Competence,* Philadelphia, University of Philadelphia Press.

HYMES, D., (1974), *Foundations in Sociolinguistics*, Philadelphia, University of Philadelphia Press.

HYMES, D., (1977), 'Towards Ethnographies of Communication', in MAYBIN, J., (ed.), (1994), *Language and Literacy in Social Practice*, Clevedon, Multilingual Matters, Ltd.

INTERNATIONAL BACCALAUREATE ORGANISATION, (1996a) *The Diploma Programme: Vade Mecum* (in English and French versions), Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (1996b), *The Diploma Programme: Group 2 Languages* (in English and French versions), Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (1997a), *Guide to the Diploma Programme* (in English and French versions), Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (1997b), *Language-specific Annexe to the Language B Guide*, Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2000a), *Subject Reports – May 2000, (Group 2 Languages: French Language B)*, Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2000b), *Subject Reports – November 2000, (Group 2 Languages: French Language B)*, Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2000c), *Teacher Observers: Grade Award Meetings*, Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2001a), *Examiners' Manual: (French Language Version)*, Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2001b), *General Instructions for May/November 2002, for Examination Paper Production,* Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2001c), *Paper Specific Instructions, Language B, Higher and Standard Levels,* Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2001d), *General Instructions for the Moderation of the Internal Assessment Component,* (French version), Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2001e), *The **IBO**: Education for Life* (Brochure in English and French versions), Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2001f), *Subject Report – May 2001, (Group 2 Languages: French Language B, in French language version)*, Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2001g), *IB World - August 2001, Issue No. 28,* Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2002a), *Guide to the Programme: Languages A2* (in English and French versions), Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2002b), *Guide to the Programme: Languages B* (in English and French versions), Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2002c), *Guide to the Programme: Languages Ab Initio* (in English and French versions), Geneva, International Baccalaureate Organisation.

INTERNATIONAL BACCALAUREATE ORGANISATION, (2002d), *Examiners' Manual: (French Language Version)*, Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, (unpublished draft), *Standardiser's Guidelines and Checklist*, Cardiff, International Baccalaureate Curriculum and Assessment Centre.

INTERNATIONAL BACCALAUREATE ORGANISATION, *Internet Website* at www.ibo.org

ISRAEL, J. B., (2000), *Authenticity and Modern Foreign Language Learning: The problem of posing authentic questions and of devising and applying criteria for effective assessment in the examinations of the International Baccalaureate for Modern Foreign Languages,* unpublished project report presented to the Open University in partial completion of the requirements for the programme leading to the *Doctorate of Education, (Ed. D.), Stage 1,* Milton Keynes, Open University.

JENKINS, R., (1992), *Pierre Bourdieu,* London, Routledge.

KRASHEN, S., (1981), *Principles and Practice in Second Language Acquisition,* London, Prentice-Hall International (UK) Ltd.

LADO, R., (1961), *Language Testing: the Construction and Use of Foreign Language Tests,* London, Longman.

LANTOLF, J. P., (2000), *Sociocultural Theory and Second Language Learning,* Oxford, Oxford University Press.

LAVE, J. and WENGER, E., (1991), 'Situated Learning: Legitimate Peripheral Participation', in MURPHY, P., (ed.), (1999), *Learners, Learning and Assessment*, London, Paul Chapman Publishing, Ltd.

LEWKOWICZ, J.A., (2000), 'Authenticity in language Testing: some outstanding questions', in ALDERSON, J.C. and BACHMAN, L.F. (eds.), *Language Testing,* (Jan. 2000, Vol 17, No. 1), London, Arnold.

LINN, R., (1993), 'Educational Assessment: Expanded Expectations and Challenges', in *Educational Evaluation and Policy Analysis*, Volume 15, Chapter 1.

LINN, R., DUNBAR, S., *et al.,* (1991), 'Complex, Performance-Based Assessment: Expectations and Validation Criteria', in *Educational Researcher*, Volume 20, Chapter 8.

McNAMARA, T., (1996), *Measuring Second Language Performance*, Harlow, Addison Wesley Longman Ltd.

McNAMARA, T., (2000), *Language Testing*, Oxford, Oxford University Press.

MILLS, J., (1997), 'The False Dasein: from Heidegger to Sartre and Psychoanalysis', in *Journal of Phenomenological Psychology*, Volume 28(1), pages 42 - 65.

OLLER, J., (1979), *Language Tests at School,* London, Longman.

OPEN UNIVERSITY, (1999), *E 836: Learning, Curriculum and Assessment,* Milton Keynes, The Open University.

ORWIG, C.J., (1999), *Ways to Approach Language Learning*, Dallas, SIL International.

OXFORD CAMBRIDGE AND RSA, (2003), *OCR Specification Synopses for GCSE*, Cambridge, University of Cambridge Local Examinations Syndicate.

QUENEAU, R., (1947), *Exercices de Style,* Paris, Gallimard.

ROGOFF, B., (1999), Audio-taped discussion with Patricia Murphy, in MURPHY, P., *et al,* (eds.), *E 836: Learning, Curriculum and Assessment: Study Guide*, Milton Keynes, The Open University.

SANDERSON, P., (1997), 'Culture and 'Subjectivity' in the Discourse of Assessment: a Case Study', in CULLINGFORD, C., (1997), *Assessment versus Evaluation,* London, Cassell.

SARTRE, J-P., (1946a), *L'Être et le Néant,* Paris, Gallimard.

SARTRE, J-P., (1946b), *L'Existentialisme est un Humanisme*, Paris, Nagel.

SARTRE, J-P., (1960), *Critique de la Raison Dialectique,* Paris, Gallimard.

SCOTT, A. (2001), *Personal Communications to the Present Researcher from the **IBO** Chief Examiner for Group 2 Languages: French,* (unpublished).

SCOTT, D.A. and USHER, R., (1996), *Understanding Educational Research,* London, Routledge.

SHAVELSON, R, *et al.,* (1992), 'Performance Assessments: Political Rhetoric and Measurement Reality', in *Educational Researcher,* Volume 21, Chapter 4.

VAN EK, J.A., (1975), 'The Threshold Level in a European Unit/Credit System for Modern Language Learning by Adults', in *Systems Development in Adult Language Learning*, Strasbourg, Council of Europe.

VAN EK, J.A., (1976), *The Threshold Level for Modern Language Learning in Schools.* Strasbourg, Council of Europe.

VAN EK, J. A. and TRIM, J.L.M., (1991), *Waystage 1990,* Strasbourg, Council of Europe.

VAN EK, J.A. and TRIM, J.L.M., (1991) *Threshold 1990*, Strasbourg, Council of Europe.

VAN EK, J.A., and TRIM, J.L.M., (1996), *Vantage Level,* Strasbourg, Council for Cultural Co-operation, Council of Europe.

VAN LIER, L., (1996), *Interaction in the Language Curriculum: Awareness, Autonomy and Authenticity,* London and New York, Longman.

WIDDOWSON, H., (1978), *Teaching Language as Communication,* Oxford, Oxford University Press.

WOOD, D., (1988), *How Children Think and Learn,* Oxford, Blackwell.

APPENDICES

APPENDIX 1

THE *INTERNATIONAL BACCALAUREATE ORGANISATION*

In its constitutional and legal status, the **IBO** is a chartered, non-profit, educational foundation, based in Geneva, Switzerland, and operating under the Swiss civil code, but with a *Curriculum and Assessment Centre* in Cardiff, South Wales, and subsidiary, regional offices throughout the rest of the world. Since 1968 it has been recognised as a non-governmental organisation by the **Council of Europe**, with consultative status in the **United Nations Educational, Scientific and Cultural Organisation, (UNESCO)**, as well as within the **United Nations Organisation** itself. It has been actively supported by **UNESCO**, the **Ford Foundation** and other international, educational funding bodies. Representatives from governments and authorised schools, as well as recognised experts in the field of education attend certain committee meetings for governing the **IBO**.

Its *Mission Statement*, adopted in 1996, declares that:

> "Through comprehensive and balanced curricula coupled with challenging assessments, the **International Baccalaureate Organisation** aims to assist schools in their endeavours to develop the individual talents of young people and teach them to relate the experience of the classroom to the realities of the world outside. Beyond intellectual rigour and high academic standards, strong emphasis is placed on ideals of international

> understanding and responsible citizenship, to the end that **IB** students may become critical and compassionate thinkers, lifelong learners and informed participants in local and world affairs, conscious of the shared humanity that binds all people together and attitudes that make for the richness of life."

By autumn 2003, this rapidly growing organisation had authorised teaching and assessment programmes in 1,493 schools scattered across more than 100 different countries of the globe. These schools are in the main, divided between private, often international schools, and state schools attached to a wide range of national institutions. In part, they fund the organisation in its activities.

The **IBO** currently devises and offers to educational institutions throughout the world, a range of three interlinking, learning programmes for children and adolescents, between the ages of 3 and 19 years. These are respectively: the *Primary Years' Programme (PYP)* for 3 to 12 year olds; the *Middle Years' Programme (MYP)* for 11 to 16 year olds; and the *Diploma Programme* for 16 to 19 year olds. It is within this latter programme, the longest-established of the three, and the largest in terms of numbers of participant schools, that the research is based.

APPENDIX 2

THE MODERATION AND EVALUATION PROCEDURES OF THE IBO FOR *GROUP 2 LANGUAGES, LANGUAGE B*

Moderation and Evaluation

Given the lack of published documentation on **IBCA** procedure for the moderation and evaluation of examination work and the assessments of *Assistant Examiners* at *Grade Award Meetings,* the research for this appendix is based on data presented in two *Reports on Attendance at Grade Award Meetings.* The first was completed for *French, Language B*, and presented to the **IBO** as a report on observation of the November 2000 examining session. It was accepted as an accurate and comprehensive record. The second was similar, but related to *German Language B* for the May 2001 examining session. Likewise, this was accepted as accurate and comprehensive. Together the two reports represent the outcomes of an analysis of data collected and reported to **IBCA** by the researcher in fulfilment of the duties of *Teacher-Observer* at the relevant meetings. As such, they are intended for use as aids in identifying conceptual frameworks, either explicit or implicit, by which understandings of criteria for assessment, moderation and evaluation, and the procedures for their application in practice are governed at **IBCA**.

For this purpose, data were progressively compressed into a summary document, care being taken to ensure that as little relevant detail as possible was lost in the process. This compressed model served as a comparator for describing, analysing and discussing the understandings and procedures adopted for the June 2001 meeting for moderating and evaluating candidate work in *German, Language B.* An element of generalisation, albeit limited, for establishing typical **IBCA** procedure was thereby validated.

The structure of the process may be summarised in stages as indicated. The data collected from *Internal Assessment* are not considered at this point, since *Grade Award Meetings* are devoted solely to the consideration of point-in-time performances by candidates in external examinations.

Stage 1: The introductory delineation of the context of the meeting with a statement of its agenda and official purposes

The aims of *Grade Award Meetings* are reiterated, as follows:

- to consider teacher and examiner comment for the previous session of examinations;
- to review the procedures and outcomes of these examinations;
- to assess the statistical information derived by **IBCA** from the relevant session, prior to the opening of the meeting;
- to reconsider and evaluate a selection of candidate work;
- thence to establish relevant grade boundaries at three points, relating to an evaluation expressed in numerical scores, and derived from the published assessment criteria;
- mathematically to calculate the remaining grade boundaries required for the application of the **IBO**'s evaluation criteria;

- to apply the grade evaluations thus derived, across the entire population of candidates at this examination.

Stage 2: Preliminary Observations

These concern the statistics for candidate entry to the examination at its differing levels; a brief description of the examining centres involved (with comments on geographical location, length of experience as a registered centre, and known peculiarities), a longitudinal referencing and a general, preliminary evaluation of the examination paper and of overall standards attained, as formally communicated by *Assistant Examiners* through the relevant *Subject Reports*.

Stage 3: Initial Procedures and Discussion

These concern statistical comparisons made by **IBCA**, of a selection of longitudinally-produced data relating to similar examinations over time, and focusing on mark distributions amongst examination candidates, per individual examination paper. They partly presume effectiveness in the prior operations of standardisation, since no final *Subject Reports* for previous sessions are taken into consideration. The criterion-referenced design of assessment, moderation and evaluation is deemed to obviate any formal need for longitudinal comparisons, establishing time and context independent norms[1]. Irregularities, uncertainties and difficulties of interpretation may be observed and discussed. Peaks and troughs evident in the graphical representation of mark-distribution by histogram are located and commented. Such information serves to highlight the relative difficulty of particular examinations and establishes a preliminary, triangulating perspective on candidate performance. Inadequacies in the design of the current examination papers are identified for subsequent reporting, with a focus on cases that, in the *Chief Examiners'* view, show evident difficulties in the understanding of candidates, attributable to the design of the examination, its rubrics, content, genre requirements and tasks.

Stage 4: The Moderation of Individual Components of the Examination by Paper (with Examiner Discussion of Problems):

Candidate scripts for *Standard Level, Paper 2* are considered first, with a broad selection of a statistically-representative sample, made by **IBCA** personnel prior to the holding of a *Grade Award Meeting*. This preliminary sampling serves roughly to determine the boundary between Grades 3 and 4. Predicted grades received from teachers are taken into consideration in the selection of copies for review, with procedure focusing initially on perceived, boundary-level candidates, then continuing with consideration of candidates more clearly attaining Grade 4, and finally correlating with candidates more clearly attaining Grade 3. Special attention is later given to candidates deemed 'at risk': that is, those for whom either a two-grade spread had been identified in the comparison of the *Assistant Examiner's* grading and the predicted grading of the teacher, or whose work was evaluated as close to a grade boundary.

The assessment criteria descriptors are re-consulted in detail and the relative popularity of tasks set established. The general evaluation criteria for the quality of language and communication, as published in the *Subject Report Guides* for previous sessions of the examinations of the relevant programme, are re-read aloud to all attending the meeting, with significant meanings emphasised. A tentative 3/4 grade boundary is proposed for a given score: that is, the process involves generalising from the basis of the specific assessment criteria as applied by the *Assistant Examiners*, and framed by the overall evaluation criteria provided in the *Subject Report Guides*[2].

After silent rereading of specific scripts by *Chief Examiners*, with *ad hoc* perusal by the *Teacher-Observer* present, uncontentious examples of Grades 4 and 3 are identified. A pre-selection of problem cases is identified at **IBCA**, numbering between 15 and 20 scripts. These are subsequently discussed, with some oral reading by examiners, as an aid to comprehension.

Extra copies of candidate work are considered until the *Chief Examiners* are satisfied, through repeated consultation of all the relevant, borderline *Assessment Criteria* descriptors, that a clear 3/4 qualitative boundary across all descriptors, has been consensually established. It is evident that the process relies significantly on the experience of *Examiners* and their understandings of fine detail in meanings and standards as employed by the **IBO**, developed over lengthy acquaintance with the programme.

Subsequent to this, a similar procedure is followed to establish the boundaries between work meriting Grades 6 and 7. The distinctions in the general evaluation criteria, as published in the examination *Subject Reports* for each language and level, are reiterated for cases lying between the 3/4 and the 6/7 boundaries. In the case of *Paper 2: Written Production*, for example, the higher grade is taken to indicate a deeper consideration of readership expectations by candidates. Such expectations may be understood from the criteria, as relating to elements of structure, appropriate presentation, originality and the concept of 'convincingness' in written production[3].

This stage of the process is completed with a review of scripts on the 3/4 boundary, in order to place certain, atypical problem cases in the 'at risk' category, for renewed review, after consideration of all the other parts of the examination[4]. Before finalising the decision of a precise mark that represents the Grade 3/4 boundary, further examples of candidate work are considered, so as to confirm, or modify this judgement. Subsequently, the same procedures are re-enacted for establishing the Grade 2/3 boundary.

The remaining grade boundaries are then interpolated mathematically, to give constant divisions between the boundaries determined through the moderating process. Further 'at risk' candidates are identified in this way, and their scripts isolated for supplementary review. The whole process thus requires the consideration of approximately 60 copies of candidate work in the two days of the meeting, normally devoted to *Language B*.

After determination of boundaries for a particular examination, longitudinal comparison is made with the boundaries established for earlier examinations.

This further stage needs noting. However, the observational data collected have been excluded from reporting in the body of this appendix, since they concern aspects of the particular examination for *Paper 1 Text Handling*. Consideration of data from this source has been included within the research design only insofar as the context of a given text relates to one of the tasks set in *Paper 2, Written Production*.

Difficulties of assessment encountered in this respect, are not considered having been analysed, discussed and resolved as far as possible, earlier in the present thesis.

Stage 5: The Moderation and Evaluation of Candidates deemed 'At Risk'

Should the further moderation and evaluation of such copies remain problematic after repetition of the given procedures, they are remarked once again, in their entirety, by a further examiner.

Stage 6: Appeals

The procedure for appeals for review of finalised results is contained in *Section B.8* of the *IB Diploma Vade Mecum*[5] and these are applied either to requests from candidates via their schools, either for a total remarking of the relevant papers, or for a report on performance, prepared by the **IBO**[6].

Final Evaluation

For convenience in reading, the data of observations are reiterated. It should be noted that in a separate operation at **IBCA**, independent of the *Grade Award Meeting*, all component grades are aggregated according to their weighted value in the relevant scheme, in a straightforward mathematical exercise. These should relate to the *General Grade Descriptors*, published together with the *Conversion Tables* for establishing grade boundaries for each discrete component of the programme, in the final *Subject Report*. This biannual publication is significant for comment that emerges on the moderation and evaluation processes. Indeed, it should be noted that the *General Grade Descriptors* are only made available to the wider public in this way. The descriptors comprise tables for converting the percentage obtained through aggregating all components of the examination, into a final **IBO** grade. (The grading system, on a seven-point scale, is common to all Groups and subject areas of the *Diploma Programme*). Also provided are the tables for converting the individual scores attributed to each section of the formal assessments by component, into a similar 'final grade' on the same seven-point scale. It may also be noted that the relationship between these various tables, at first glance apparently reworking similar data, is not made explicit in the documentation published[7]. Their significance lies as a control for distortion in the final aggregation of component scores and grades, and as such is discussed in the conclusions of Chapter 7 of the thesis.

APPENDIX 3

THE RESEARCH INSTRUMENT

Conditions for Authentication:

Assessment Criteria Developed from the Work of Van Lier For the Identification of Features of Authentic Language Use

EVIDENCE FOR CURRICULAR AUTHENTICATION: a focus on 'self'	CRITERIA	None	Little	Adequate	Significant	Abundant
Creator authenticity and the notion of 'self'	**The oral and written production displays autonomous treatment of the task set.** *The initial focus of this criterion is on authenticity as a recognition of the foundational primacy of the existence of 'self' as the starting point of all conscious thought. In communicative activity, this requires the expression of 'self' as the originator of communication, through free choice of focus of individual attention, even though the operation of this choice falls within the parameters of the social context for communication that has been defined in the task set.*	The text fails to establish relevance either to the task set or for the typical audience addressed. *No answer has been indicated for the question: "Why speak or write?"* The creator of the text fails to inform the audience of relevant, personal concerns in communicating in this context. *Why communicate?* The text fails to establish any sense of 'voice' for its creator. *It is completely unclear as to who has spoken or written.* The meaning of the text as a message is confused.	The text shows a discernible relationship to the requirements of the task set, in terms of personal response. *Some sort of answer has been indicated for the question: "Why speak or write?"* The creator of the text indicates only vaguely, the personal concerns that have motivated communication. *Why communicate?* The text supplies evidence of an original, authorial 'voice', though perhaps fleetingly and incoherently. *It is vaguely perceivable who has spoken or written.* A message can be	The text evidently relates to the explicit requirements of the task set, in terms of setting a personal response. *An unequivocal, if simple answer has been indicated for the question: "Why speak or write?"* The creator of the text has clearly identifiable, personal concerns that are at least partially coherent as a set, and relevant to the audience addressed. *The question "Why communicate?" is answered in some way.* The text establishes some sense of 'personality' for its creator. *It is perceivable who,*	The text relates coherently and consistently to the requirements of the task set, in terms of a personal response. *The question: "Why speak or write?" has received a clear and consistent answer.* The creator of the text's personal reasons for communicating with this particular audience are unambiguous and coherent. *The question "Why communicate?" is clearly answered.* The text establishes a sense of 'individuality' in the personality of its creator. *It is quite evident who, as an individual, has spoken or written.*	The text relates coherently, consistently and perhaps imaginatively to the requirements of the task set, in terms of a personal response. *The question: "Why speak or write?" has received a clear, consistent and interesting answer.* The creator of the text's personal reasons for communicating with this particular audience are unambiguous, coherent, and in part sophisticated. *The question "Why communicate?" is answered in a satisfying, perhaps enlightening way.* The text establishes a sense of complex,

	This authenticity is also tied to autonomy as a notion of 'self' conceptualising and constructing an evolution of identity as a consequence of acts of free choice that engage 'self' in interaction with the environment within which it is located.	*The audience is left thinking "What is going on?"* The organisation of the communication is haphazard and chaotic. *The audience is left thinking, "I can't follow this".*	understood, even if with difficulty. *The audience is left thinking, "It is not easy to follow what is going on."*	*as an individual, has spoken or written.* A message can be understood, even if such understanding can sometimes seem ambiguous. *The audience is left thinking, "It is fairly easy to follow what is going on."*	The message is clear and coherent, creating the impression that the creator has 'something to say'. *The audience is left thinking "I want to follow what is going on and maintain this willingness to focus attention up to the end of the message given."*	sophisticated, or multifaceted individuality in the personality of its creator. *It is evident who, as an individual, has spoken or written. The manner commands attention.* The message is clear, coherent and developed with sophistication, creating the impression that what has been said or written is important and worth the investment of time in listening or reading. *The audience is left thinking "I want to follow what is going on and maintain this willingness to focus attention up to the end of the message given. I feel strongly that I have learnt something about the creator of the message as an individual."*

EVIDENCE FOR CURRICULAR AUTHENTICATION: a focus on 'self'	CRITERIA	None	Little	Adequate	Significant	Abundant
Creator Authenticity and awareness of the existence of 'other'	**The oral, or written production displays awareness of the expectations of a typical audience.** *The initial focus of this criterion is on authenticity as an awareness of the existence of 'others', and of a dialogical context in which meanings can be constructed dialectically and socially.* *This leads to an expectation that communication and intersubjective responses will continue in some form, in order to allow 'self' to develop and extend meanings in interactions with both the material context of the communication and the 'other'*	The text makes no reference to the intended audience. The creator of the text fails to indicate any response to audience requirements, as set by the task proposed. There is no evident answer to the question: *Why communicate with these people?* The creator of the text fails to indicate a context in which communication should commence and be developed, with an expectation of response from the audience as set in the task. *The question: "Why communicate*	The text shows an undeveloped awareness of the intended audience. The nature of this audience may not necessarily be precisely defined, or even coherently identified. The creator of the text responds inconsistently, or imprecisely to the requirements of a typical audience. There is an imprecise answer to the question: *Why communicate with these people?* There is little indication of the reasons for communicating in the way chosen, or of an expectation of	The text shows a clear awareness of the general nature of the audience addressed. This is displayed directly and concretely, if in simplified terms The creator of the text indicates a coherent response to the requirements of a typical audience. *The question "Why communicate with these people?" is answered simply, but unambiguously.* The reasons for creating the text at this point in time and in this manner, are evident to some degree. A purpose for communicating is indicated. An expectation of	The creator of the text shows a clear awareness of the relevant, specific qualities of the audience addressed. This is displayed through variation in perspective that shows a multi-dimensional awareness of audience interests. The creator of the text responds in a coherent and satisfying way to the explicit requirements of the audience addressed. *The question "Why communicate with these people?" is unambiguously answered, raising no thought that anything is lacking.*	The creator of the text shows an appropriate and nuanced awareness of the relevant, specific qualities of the audience addressed. This is displayed through variations in perspective that show a subtle awareness of audience interests and may nurture further interests in an 'enlightening' manner. The creator of the text evidently responds both to explicit, and to certain implicitly perceived requirements of the audience in a continually coherent manner. *The question "Why communicate with these people?" is answered with some*

	subjectivities addressed.	*at this time and in this way?" remains unanswered.*	continuation of the dialogue initiated. *The question: "Why communicate at this time and in this way?" is only partially, and perhaps ambiguously, answered.*	feedback that continues the flow of communication may be discernible. *The questions "Why communicate at this time and in this way?" are both answered simply, yet unequivocally.*	The reasons for creating the text at this point in time and in this manner are clear. The text is explicitly linked to its prompt stimulus. (The reasons for choosing this stimulus over others may be explicitly stated). The text stands as a contribution of ideas worth communicating, and concludes with an expectation, at least implicit, of an audience reaction. *The questions "Why start communicating? Why continue with this communication? And What next?" are at least implicitly answered.*	*depth of perception indicating a rounded awareness of their expectations, both explicit and implicit.* The creation of the text is clearly contextualised in time and choice of form, as an appropriate response to a defined task, as a coherent project for communication, and as an invitation to continue communication beyond the statement of the text itself, where appropriate. *The questions "Why start communicating? Why continue with this communication?" and: " What next?" are satisfyingly answered.*

EVIDENCE FOR CURRICULAR AUTHENTICATION: a focus on 'self'	CRITERIA	None	Little	Adequate	Significant	Abundant
Finder Authenticity, or the resourcefulness of the communicator in finding material for communication: Self and the focusing on, the selection of objects of attention	**The text produced displays resourcefulness in the finding of appropriate material for communication.** *The focus of this criterion is the demonstration of an understanding that 'self' is located in temporal and spatial contexts, from which meaning is constructed in interactive, communicative and dialectical relationships: such interaction is required as essential in order for the cultural contextualisation of 'self' within society to take place.*	The choice of content of the text produced is irrelevant to the task set. *No answer is discernible to the questions: "Why and from what material has this been produced?"* The organisation of the content of the text is haphazard, incoherent, and very difficult to follow. *The question "What is being said" is unanswerable.*	The choice of content of the text displays some relevance to the requirements of the task set, albeit inconsistently and with some inappropriateness, confusion or lack of precision. *A vague answer to the questions: "Why, and from what has this been produced?" is discernible.* There is some evidence of coherence in the sequence of ideas (with or without factual detail), as presented in the text. *The question "What is being said" can be answered at least in part, but consistency is lacking, and perhaps significant*	The choice of content of the text produced addresses the explicit requirements of the task set, with little irrelevant detail or comment. *An answer to the questions: "Why, and from what has this been produced?" has been supplied, albeit not perhaps entirely meeting the expectations of the audience for clarity.* The ideas and factual detail presented in the text are coherently and appropriately sequenced, although this may not necessarily be in a fully consistent and convincing fashion. *The question "What is being said" can be*	The choice of content of the text produced openly addresses the explicit requirements of the task set, without any significant moments of irrelevance. *An unambiguous answer to the questions: "Why, and from what has this been produced?" has been supplied, with the audience left feeling satisfied that the task has been completed in an appropriate and rounded fashion.* The ideas and factual detail presented are appropriately varied in content. *The audience judges the text as nuanced, balanced and convincing, even if*	The choice of content of the text produced addresses the requirements of the task set with sophistication and resourcefulness: that is, with some evidence of an awareness of both explicit and implicit expectations on the part of a typical audience, and perhaps with original insight. *An answer to the questions: "Why, and from what has this been produced?" has been supplied, with the audience left feeling that the task has been completed in a stimulating and fully rounded fashion.* The ideas presented are rich and varied,

			omissions, contrary to the typical audience's expectations, remain. The whole is unconvincing as an argument.	*answered, but consistency may be lacking, and perhaps certain omissions, contrary to the typical audience's expectations, remain. The arguments are acceptable, albeit with reservations.*	*further discussion may be appropriate.* The sequencing of ideas and facts presented is both coherent and consistent. *The text flows as a sequence of thought, without hiccough from start to finish.*	supported with relevant detail that refers to a world beyond the self (facts, or the opinions of others): that is, evidence is provided of a comprehensive range of argument and an imaginative response to the requirements of the task. *The audience judges the text as satisfying, nuanced and balanced, possibly stimulating the desire for further discussion or comment.* The sequencing of ideas and facts presented is coherent and consistent, stimulating the interest of the audience to persevere in focusing attention on the text as it progresses in time. *The text flows as a sequence of thought, without hiccough from start to finish, encouraging the audience to continue to follow what is being communicated.*

EVIDENCE FOR CURRICULAR AUTHENTICATION: a focus on 'self'	***CRITERIA***	***None***	***Little***	***Adequate***	***Significant***	***Abundant***
User Authenticity, or the recognition by self of other, in the forms of linkages with socio-cultural tradition and convention in order to allow an initiation of communication	**The text produced is used in a manner appropriate to the task set.** *The focus is on the examination candidate's recognition of the forms by which communication in social contexts can be mediated, in order to facilitate communication from self, to other. This requires an acknowledgement, be it implicit or explicit, of appropriate socio-cultural traditions and conventions, although these may not necessarily determine the shape or content of the language produced, or the genre format chosen.*	The construction and presentation of the text bears no relation to the traditions and conventions of 'genre', appropriate to the task set and as expected by any typical audience. No rationale for deviating from such conventions is given. *The text produced appears 'foreign' to the audience addressed, and meets no pre-existing expectations.* The language employed displays no features that typify the traditions and conventions of the 'genre' chosen, and as expected by a typical, but tolerant, audience. No	The construction and presentation of the text shows features, albeit perhaps inconsistently, of the traditions and conventions of 'genre', appropriate to the task set and as expected by any typical audience. Deviations from such conventions are recognised as such. *The text produced appears 'foreign', but is in part, accessible to the audience addressed. It meets certain expectations, without requiring knowledge of the traditions and culture of the producer of the text, on the part of the audience.* The language	The construction and presentation of the text shows clear relationships, albeit perhaps occasionally inconsistently, with the traditions and conventions of 'genre', appropriate to the task set and as expected by any typical audience. A rationale is given for any deviations from such convention. *The text produced may appear partly 'foreign', but is wholly accessible to the audience addressed. It meets certain expectations, without requiring any special knowledge of the traditions and culture of the producer of the text, on the part of the audience.*	The construction and presentation of the text shows clear, consistent relationships with the traditions and conventions of 'genre', appropriate to the task set and as expected by any typical audience. Deviations from such convention are acknowledged and inventive, if not always with full appropriateness. *The text produced is wholly accessible to the audience addressed. It meets certain expectations, without requiring any special knowledge of the traditions and culture of the producer of the text, on the part of the*	The construction and presentation of the text shows clear recognition of, and respect for the traditions and conventions of 'genre', appropriate to the task set, with evidence of an ability to extend traditional values and develop conventional usage with imaginative resourcefulness and perhaps some originality. A clear and convincing rationale is given for any deviations from such convention. *The text produced is wholly accessible to the audience addressed. It meets all expectations, without requiring any special knowledge of*

		rationale for deviating from such conventions is given *The text appears wholly foreign in its language, perhaps as if it has been literally translated, word-for-word, from the mother tongue in which it has been structured and constructed.*	employed displays some features, be they explicit or implicit, that typify the traditions and conventions of the 'genre' chosen, and as expected by a typical, but tolerant, audience. Deviations from such conventions are recognised as such. *The text appears rather 'foreign' in its language, as if the author's thoughts had been constructed and structured in another tongue.*	The language employed displays clear features that typify the traditions and conventions of the 'genre' chosen, as expected by a typical, but tolerant, audience, and partly explicitly so. A rationale is given for any deviations from such convention. *The text appears 'normal'. If perhaps inconsistent in its language, as if the author's thoughts had been constructed and structured in the language of the audience, albeit perhaps inexpertly so.*	*audience.* The language employed displays with consistency, the traditions and conventions of the 'genre' chosen, as expected by a typical audience, both implicitly and explicitly so. Deviations from such convention are acknowledged and inventive, if not always with full appropriateness. *The text appears mostly 'normal' in its language, as if the author's thoughts had been fluently constructed and structured in the language of the audience.*	*the traditions and culture of the producer of the text, on the part of the audience.* The language employed displays with resourceful variety, and perhaps imaginative usage the traditions and conventions of the 'genre' chosen, as expected by a demanding audience, both implicitly and explicitly so. A rationale is given for any deviations from such convention. *The text appears 'rich' and 'fluent' in its language, as if the author's thoughts had been expertly constructed and structured in the language of the audience.*

EVIDENCE FOR PRAGMATIC AUTHENTICATION: The creation of 'realism'	***CRITERIA***	***None***	***Little***	***Adequate***	***Significant***	***Abundant***
Authenticity of Context: the indication of willingness to share cultural perspectives through the possibilities afforded in initiating communication and in recognition of, and respect for, the traditions and conventions of other cultures.	**The choice of genre and language used displays specific features typifying the commonly-understood setting of the task.** *The focus is on the examination candidate's reproduction of the forms by which 'typical' communication in social contexts are formulated, in order to regulate communication from self, to other. This requires an acknowledgement, be it implicit or explicit, of appropriate genre formats and choice of language, although these may not necessarily exclude features of originality of presentation and*	The construction and presentation of the text bears no relation to the traditions and conventions of 'genre', appropriate to the task set and as expected by any typical audience. No rationale for deviating from such conventions is given. *The text produced appears 'atypical' to the audience addressed, and meets no pre-existing expectations. A sense of no compromise with tradition for the sake of ease of communication is created.* The language employed displays no features that typify the traditions and conventions of the	The construction and presentation of the text shows features, albeit perhaps inconsistently, of the traditions and conventions of 'genre', appropriate to the task set and as expected by any typical audience. Deviations from such conventions are recognised as such. *The text produced appears 'atypical', but is in part, likely to be recognised as a 'realistic' representation by the audience addressed. It meets certain expectations, by partly conforming to the cultural norms of the society of the language used. It requires however, some knowledge of the traditions and*	The construction and presentation of the text shows clear relationships, albeit perhaps with occasional inconsistency, with the traditions and conventions of 'genre', appropriate to the task set and as expected by any typical audience. A rationale is given for any deviations from such convention. *The text produced may appear partly 'unusual', but is recognised as a 'realistic' representation and is wholly accessible to the audience addressed. It meets certain expectations, by conforming to the cultural norms of the society of the language used. It*	The construction and presentation of the text shows clear, consistent relationships with the traditions and conventions of 'genre', appropriate to the task set and as expected by any typical audience. Deviations from such convention are acknowledged and inventive, if not always with full appropriateness. *The text produced is wholly accessible to the audience addressed and recognised as 'authentic'. It meets expectations, by conforming to the cultural norms of the society of the language used, though may contain original*	The construction and presentation of the text shows clear recognition of, and respect for the traditions and conventions of 'genre', appropriate to the task set, with evidence of an ability to extend traditional values and develop conventional usage with imaginative resourcefulness and perhaps some originality. A convincing rationale is evident, be it implicit or explicit, for any deviations from such convention. *The text produced is wholly accessible to the audience addressed. It meets all expectations, and appears as fully 'authentic', without requiring any special*

	expression.	'genre' chosen, and as expected by a typical, but tolerant, audience. No rationale for deviating from such conventions is given *The text appears wholly 'atypical' in its language, which is decontextualised. The creator of the text shows no evidence of familiarity with the norms of the culture addressed.*	*culture of the producer of the text, on the part of the audience, in order to be better understood. As a result, it may create a sense of confusion and incoherence due to inadequate knowledge of the culture addressed.* The language employed displays some features, be they explicit or implicit, that typify the traditions and conventions of the 'genre' chosen, and as expected by a typical, but tolerant, audience. Deviations from such conventions are in some way, perhaps implicitly, recognised as such. *The text appears rather 'atypical' in its language, and requires nonetheless, some knowledge of the language of the producer of the text, on the part of the audience, in order to be better understood.*	*requires no special knowledge of the traditions and culture of the producer of the text, on the part of the audience.* The language employed displays clear features that typify the traditions and conventions of the 'genre' chosen, as expected by a typical, but tolerant, audience, and partly explicitly so. A rationale is given for any deviations from such convention. *The text appears 'normal', if perhaps occasionally unusual in its use of language. It requires little knowledge of the language of the producer of the text, on the part of the audience, in order to be clearly understood. However, the language used may occasionally confuse and create a momentary sense of incoherence due to small lapses in*	*developments, even if these are not wholly successful in their effect. No knowledge of the traditions and culture of the producer of the text, on the part of the audience, is required for complete comprehension.* The language employed accords with the traditions and conventions of the 'genre' chosen, as expected by a typical audience, both implicitly and explicitly so. Deviations from such convention are acknowledged and inventive, if not always with full appropriateness. *The text appears 'normal' in its language. It requires no knowledge of the language of the producer of the text, on the part of the audience, in order to be clearly understood.*	*knowledge of the traditions and culture of the producer of the text, on the part of the audience. In certain aspects, it may appear 'creative'.* The language employed displays with resourceful variety, and perhaps imaginative usage, a sympathetic response to the traditions and conventions of the 'genre' chosen, as expected by a demanding audience, both implicitly and explicitly so. A rationale is evident for any deviations from such convention, which may be 'extended' by the example of the text created. *The text appears 'rich', 'fluent' and perhaps occasionally 'inventive' in its use of language. It requires no knowledge of the language of the producer of the text, on the part of the audience, in order to be fully understood.*

			As a result, the language used may create a sense of confusion and incoherence due to inadequate knowledge of the structures and lexis required in this context.	*knowledge of the structures and lexis required in this context.*		

EVIDENCE FOR CURRICULAR AUTHENTICATION: a focus on 'self'	CRITERIA	None	Little	Adequate	Significant	Abundant
Authenticity of Purpose: issues of transparency and self-awareness in the choice of genre for expression and message for communication: the favouring of some form of change in the perspectives and knowledge of the audience of the text created.	**The text produced, whether oral or written, displays transparency of intended effects on its audience, and may propose an explicit outcome as a consequence of reception. It shows awareness that listening and reading produce change in the audience and seeks to mould such change by intentionality, as expressed in the choice of form and message.** *The focus is on the construction of further communicative acts by proposing a self-determined, or sanctioned agenda, for consideration by others who will in turn,*	The authorial purpose of communication is concealed, obscured or lacking. The author may be unaware of any purpose for proposing the communication to be assessed. *The text provides no indication of any rationale, either implicit or explicit. It seeks to create no linkages with its audience, and stands isolated as an artefact. It creates no evidence of reflexive meaning for the 'self' that composed it.*	The authorial purpose of communication is partially concealed, obscured or lacking. It is difficult, but possible, to discern, involving an act of intention to interpret on the part of its audience. The author may be unaware of any explicit purpose for proposing the communication to be assessed. The purpose of the text remains ambiguous. *The text provides little indication of any explicit rationale, though implicit purposes for communication may be discernible. It seeks to create few, and merely ambiguous linkages with its audience, and may stand in isolation*	An authorial purpose of communication is supplied and discernible, either as explicit statement or as implicit intention. It may involve an act of intention to interpret on the part of its audience. The author may be insufficiently aware of explicit purposes for proposing the communication to be assessed, but within the context of the culture of the target language's users, intentions are unambiguously apprehended. The overall purpose of the text may remain inconsistent, but a purpose is understood. *The text may provide an explicit rationale,*	An authorial purpose of communication is supplied and unambiguously clear, either as explicit statement or as implicit intention. It may however seek an act of intention to interpret on the part of its audience. The author shows awareness of explicit purposes for proposing the communication to be assessed, and within the context of the culture of the target language's users, intentions are unambiguous, and easy to apprehend. The overall purpose of the text relates to the choice of form and content, with clear purposes understood.	An authorial purpose of communication is supplied and unambiguously clear, either as explicit statement or as implicit intention. It may however seek an act of intention to interpret on the part of its audience. The author shows awareness of explicit purposes for proposing the communication to be assessed, and within the context of the culture of the target language's users, intentions are unambiguous, and easy to apprehend. The overall purpose of the text is consistent and well integrated with the choice of form and content. All purposes appear as appropriate

	react to, and so develop communication from the influence thus proposed.		*as a barely meaningful artefact. It creates little evidence of reflexive meaning for the 'self' that composed it.*	*though implicit purposes for communication are easy to discern for a discriminating audience. It seeks to create linkages with this audience, though may stand in isolation as an artefact in its own right. It suggests an element of reflexive meaning for the 'self' that composed it.*	*The text may provide an explicit rationale, though implicit purposes for communication are easy to discern for an appropriate audience, as defined in the task set. It seeks to create linkages with this audiënce, though may stand in isolation as an artefact in its own right. It proposes an element of reflexive meaning for the 'self' that composed it.*	to the task set. *The text may provide an explicit rationale, though implicit purposes for communication are easy to discern for an appropriate audience, as defined in the task set. It seeks to create linkages with this audience, though may stand in harmonious isolation as an artefact in its own right. It proposes a clear status of reflexive meaning for the 'self' that composed it.*

EVIDENCE FOR CURRICULAR AUTHENTICATION: a focus on 'self'	CRITERIA	None	Little	Adequate	Significant	Abundant
Authenticity of interaction: the recognition of issues of power as a significant factor in determining the quality of social relationships: questions of balance, 'convincingness' and validity	**The text produced, whether oral or written, displays 'symmetry' in the participation rights and duties of linguistic interchange, standing as evidence of 'autonomy' for the 'self', in intentional interaction via language, with 'others' recognised as equally autonomous.** *The focus is on the awareness of needs for compromise in the choice of form, message and expression, for the sake of effective communication with the designated audience. Such compromise allows the successful*	The text seems 'unbalanced', with either the concerns of 'self', or the perceived requirements of the audience identified predominant, the one to the exclusion of the other. *The text fails to 'convince' the listener or reader, by being exclusively concerned with a personal and wholly subjective agenda, and/or by being framed in a bizarrely idiosyncratic form. Alternatively, the text fails to 'convince' by slavishly following and reproducing received ideas and genres to the effect on the audience that the author has "nothing to say", and*	The balance of views, ideas and forms chosen for the text seems 'not right', although evidence is present of an awareness of the likely perspectives of the audience, and of a willingness to express 'self' in interaction with this audience. *The text fails largely to 'convince' the listener or reader, by being overly concerned with a personal and mainly subjective agenda, and/or by being framed in an idiosyncratic form that is not easy to follow. Alternatively, the text fails unambiguously to 'convince' by unimaginatively following and reproducing received*	A balance of views, ideas and forms chosen for the text is established, though may appear at times to be 'not quite right'. Evidence is present of an awareness of the likely perspectives of the audience, and of a willingness to express 'self' in interaction with this audience. *The text may 'convince' the listener or reader, as being 'competent. That is, the choice of form and content responds to the perceived requirements of the task in a reasonably balanced way. A personal and subjective agenda is provided, and is framed in a conventional form.*	A good balance of views, ideas and forms chosen for the text is well established. The whole appears to be 'right'. Evidence is present, either explicitly or implicitly so, of an awareness of the likely perspectives of the audience, and of a ready willingness to express 'self' in interaction with this audience. *The text 'convinces' the tolerant listener or reader, as being 'competent. That is, the choice of form and content responds to the perceived requirements of the task in a satisfyingly balanced way. A personal and subjective agenda is*	A good balance of views, ideas and forms chosen for the text is persuasively established. The whole appears to be 'right'. Evidence is present, either explicitly or implicitly so, of insight into the likely perspectives of the audience, and of an eager willingness to express 'self' in interaction with this audience. *The text 'convinces' the demanding, or sceptical listener or reader, as being very 'competent'. That is, the choice of form and content responds imaginatively to the perceived requirements of the task in a satisfyingly balanced way. A personal and*

	integration of 'self' within the context of the society of the audience proposed. It avoids either: determination of 'self' by the agenda of others, slavishly followed; or the determination of 'other' by the authoritarian dictates of the 'self' proposed.	appears unwilling to provide occasion for interaction and further dialogue.	ideas and genres to the effect on the audience that the author has "little to say", and appears reluctant to provide substantive occasion for interaction and further dialogue. The text leaves doubts as to the validity of its form and content in the mind of the audience.	Idiosyncratic elements are easy to follow. Alternatively, the text may 'convince', although somewhat ambiguously. It may unimaginatively follow and reproduce received ideas and genres. However, in this case, the effect on the audience is that the author has "something to say". The author appears prepared to enter the occasion presented for interaction and further dialogue. The text may leave some doubt as to the validity of the whole in its form and content. However, valid aspects are present in the mind of the audience, even if only in the consideration of details.	provided, and is framed in a harmonious form. Idiosyncratic elements are easy to follow. Alternatively, the text 'convinces', even though it may reproduce received ideas and genres. However, in this case, the effect on the audience is that the author has "something interesting to say". The author is prepared to enter the occasion presented for interaction and further dialogue. The text leaves little doubt as to the validity of the whole in its form and content. Detailed points support the creation of this sense of overall validity.	subjective agenda is provided, is perceived as relevant, and is framed in a harmonious form. Idiosyncratic elements are easy to follow and appear creative, or inspired. Alternatively, the text wholly 'convinces', despite reproducing some received ideas and genres. In all cases, the effect on the audience is that the author has "something interesting and important to say". The author is fully prepared to enter the occasion presented for interaction and further dialogue. The text leaves no significant doubt as to the validity of the whole in its form and content. Detailed points support the creation of this sense of overall validity and contribute to the persuasiveness of the whole.

<u>EVIDENCE FOR PERSONAL AUTHENTICATION</u>	<u>CRITERIA</u>	<u>None</u>	<u>Little</u>	<u>Adequate</u>	<u>Significant</u>	<u>Abundant</u>
<u>Existential Authenticity: the expression and social construction of a notion of 'self' through (communicative) actions</u>	**The text production displays evidence of a personal commitment to the activity set as a task, and a willingness to use language communicatively in order to extend the bounds of self, through exploration of the communicative world linking 'self' with 'other'** *The focus is on an initial recognition and expression of 'self' that through intention, engages with an environment determined by the task and the language of communication. This 'self' displays awareness that it is through the exercise of choice that the form and content of the subsequent*	The text gives no evidence of a desire on the part of the author, to communicate in any form, either with 'self' or with an audience. *The text appears to close down all avenues for communication. Its tone is negative and exclusive. A refusal to use the language of communication set by the task, or a refusal to engage with the task as set, may be evident.*	The text gives evidence of a desire on the part of the author, to communicate in some form, even if ambiguously, or partially presented, either with 'self' or with an audience. *The text appears to show an intent to communicate, even though expectations of response in any form by the audience addressed may be inchoate. Its tone may be at times negative and exclusive. However, there is no evident refusal to use the language of communication set by the task, or refusal to engage with the task as set, even though the outcome may be unsatisfactory for*	The text gives evidence of a desire on the part of the author to communicate, even if only implicitly, either with 'self', or with an audience. This desire may be shown by the use of forms and content that may attempt to elicit responses in others, or an extension of the boundaries of self through reflection and self-exploration. A personal, or social 'goal', contextualised by the task set, is at least implicit. *The text shows an intent to communicate, and influence possible responses in any form by the audience addressed. Its tone is positive and inclusive, at least*	The text gives clear evidence of a desire on the part of the author to communicate, even if only implicitly, either with 'self', or with an audience. This desire is likely to be shown by the use of forms and content that may attempt to elicit responses in others, or an extension of the boundaries of self through reflection and self-exploration. A personal, or social 'goal', contextualised by the task set, is apparent. *The text shows the intent to communicate, and attempts in part to shape possible responses from the audience addressed. Its tone is positive and inclusive, at least*	The text gives clear evidence of a desire on the part of the author to communicate, either with 'self', or with an audience. This desire is likely to be shown by the use of forms and content that attempt to elicit responses in others, or an extension of the boundaries of self through reflection and self-exploration. A personal, or social 'goal', to be achieved in an integrated, balanced way, as contextualised by the task set, is clearly communicated. *The text shows evident intent to communicate, and attempts to shape possible responses from the audience addressed. Its tone is*

	communication is shaped. Explicitly, or implicitly, it accepts responsibility for the contributions made it its name.		some or all concerned.	implicitly so. There is clear evidence of a willingness to use the language of communication set by the task, and to engage with the task as set, even though the outcome may be in some respects unsatisfactory for some or all concerned.	implicitly so. Its agenda is purposeful, perhaps imaginative and creative. There is clear evidence of a willingness to use the language of communication set by the task, and to engage with the task as set. A sense of enthusiasm for communication in the language set may underlie and reinforce the qualities noted above.	positive and inclusive, perhaps explicitly so. Its agenda is purposeful, 'original', imaginative and creative. The language of communication set by the task is used with a sense of relish, and engagement with the task set is eager. A sense of enthusiasm for communication in the language set underlies and reinforces the qualities noted above.

EVIDENCE FOR CURRICULAR AUTHENTICATION: a focus on 'self'	***CRITERIA***	***None***	***Little***	***Adequate***	***Significant***	***Abundant***
Intrinsic Authenticity: issues of self-determination through continuing processes of choice in socio-temporal contexts	**The text production communicates a sense of self-determination in the author, as if recognising that exploration and extension of 'self' occurs in part through engagement in the composition and presentation of the chosen text.** *The focus is on the recognition of possibilities for change through the exercise of choice in addressing the constraints of the task set. This choice is expressed through the selection of a focus for awareness on the part of the author, and the commitment to communicative 'action' in intentionally composing the text presented as a*	The text produced is incoherent, or lacking in communicative value, setting no agenda for reflection or discussion. *There is no evidence of focussing of awareness on the part of the author, or of attempts to manipulate the constraints of the task set. There is no evidence of any appeal to the reflective self, or to the intended audience for participation in communication. There is no sequencing of the material presented in a fashion that could illustrate development, and no framing for the selection of sections of material that could*	The text produced is confused, or lacking in anything other than elementary communicative value. It sets no clear agenda for reflection or discussion, and shows little awareness of the status of its communicative value for the intended audience. *The focus of awareness of the author may be obscure, although evidence is present of an attempt to manipulate the constraints of the task set. There is little evidence of any appeal to the reflective self, or to the intended audience for participation in communication. There is little coherent*	The text produced is clear, though perhaps lacking in anything other than elementary communicative value. It sets an agenda for reflection or discussion, and shows awareness of the status of its communicative value for the intended audience. *The focus of awareness of the author may be disparate, although the constraints of the task set have been appropriately adapted to allow for the expression of personal views and choices of material. There is evidence of an appeal either to the reflective self, or to the intended audience for participation in*	The text produced is clear, and succeeds in capturing the attention and interest of its audience. It sets an agenda for reflection or discussion, and shows awareness of the status of its communicative value for the intended audience. It achieves a 'flow' of coherence. *The focus of awareness of the author is clearly chosen. The constraints of the task set have been appropriately adapted to allow for the expression of personal views and choices of material. The text makes an unequivocal appeal either to the reflective self, or to the intended audience for*	The text produced is clear, and succeeds in capturing the attention and interest of its audience for its inventiveness, inspiration or originality. It sets an agenda for reflection or discussion, and makes a clear statement of its own communicative value for the intended audience. It achieves a satisfying 'flow' of coherence from start to finish. *The focus of awareness of the author is clearly chosen. The constraints of the task set have been imaginatively, yet coherently adapted to allow for the expression of personal views and choices of material. The text makes an unequivocal appeal*

	positive statement of 'selfhood' in development through interaction with the audience addressed. The exercise of such choice is understandable and communicable insofar as it is both internally and externally 'coherent' as a sequencing of selective focus and acts of engagement across the time taken to 'complete' the task set.	*allow such sequencing to be coherent.*	*sequencing of the material presented in a fashion that could illustrate development. The framing provided for the selection of sections of material that could allow such sequencing to be coherent is weak, and as such, inconsistency appears at various stages in the unfolding of the text for the audience intended.*	*communication. There is coherent sequencing of the material presented in a fashion that could illustrate development, although this may not always be consistent for the audience intended. The framing provided for the selection of sections of material that could allows such sequencing to be coherent is appropriate, and as such, inconsistency is mostly inconsequential in the unfolding of the text for the audience. A sense of 'direction' in the communication presented, is evident.*	*participation in communication. There is coherent, consistent sequencing of the material presented, illustrating a development of theme across the length of the text produced. The framing provided for the selection of sections of material that allows such sequencing to be coherent is appropriate, and as such, there is no inconsistency in the unfolding of the text for the audience. A sense of 'direction' in the communication presented, is clear. The text creates a sense of being confident as a piece of work, and as such is consistently convincing.*	*either to the reflective self, or to the intended audience for participation in communication. There is coherent, consistent sequencing of the material presented, illustrating a development of theme across the length of the text produced, with at least an implicit resistance to the closure of discourse on completion. The framing provided for the selection of sections of material that allows such sequencing to be coherent is satisfyingly appro-priate, and as such, there is no incon-sistency in the unfolding of the text for the audience. A sense of purposeful 'direction' in the communication presented, is clear. The text creates a sense of being 'authoritative' and confident as a piece of work, and as such is highly convincing.*

EVIDENCE FOR CURRICULAR AUTHENTICATION: a focus on 'self'	***CRITERIA***	***None***	***Little***	***Adequate***	***Significant***	***Abundant***
Autotelic Authenticity (after Csikszentmihalyi [1]) The experience and expression of 'flow' as 'optimal experience': issues of coherence and psychological balance	**The text produced displays the integration of features of both existential and intrinsic authenticity, as assessed above.** *The communicative activity undertaken illustrates the control of consciousness by self within the defined parameters of the task in its socio-historical context. That is, the focus of awareness is controlled and centred on relevant aspects of the task so as to create internal 'coherence' as a rationale, expressed either implicitly, or explicitly. The production avoids irrelevance, obsessive imbalance in self-*	The text produced appears unbalanced and incoherent as a result. The author gives the audience or interlocutors the impression of being either bored with the subject, or alienated from it. *There is no evidence of any attempt to engage with the task as set, or any desire to communicate in any form. The text proposes neither food for reflective thought by 'self', nor desire to interact with others. Attention appears unfocussed. The text lacks both evidence of autonomy on the part of the 'self' that has composed it, and*	The text produced appears unbalanced. This imbalance gives rise to some confusion or feeling of dissatisfaction on the part of the audience or interlocutor, as a result. The author gives the audience or interlocutors the impression of being only superficially engaged with the subject, as if it does not appeal, or hostile to it in itself, for inchoate reasons that have not been clearly expressed. *There is little evidence of an attempt to engage meaningfully with the task as set, though a desire to communicate is detectable. The text proposes little food for*	The text produced appears balanced, if not necessarily uniformly so. The overall effect is however satisfying for the audience or interlocutor, despite any notable imbalance. The author gives the audience or interlocutors the impression of being engaged with the subject, as if it appeals. *There is clear evidence of an attempt to engage meaningfully with the task as set, and the desire to communicate is evident. The text proposes either food for reflective thought by 'self', or a desire to*	The text produced appears satisfyingly balanced and coherent. The author gives the audience or interlocutors the impression of being engaged with the task, as if it appeals and is worth exploring through the form and content of a text production. *There is clear evidence of an attempt to engage more than superficially with the task as set, and the desire to communicate is evident. The text proposes either food for reflective thought by 'self', or a desire to interact with others, or both. Attention appears*	The text produced appears satisfyingly balanced, coherent and autonomously produced. The author gives the audience or interlocutors the impression of being deeply engaged with the task, as if it appeals and is worth exploring through the form and content of a text production. *There is clear evidence of focussed engagement with the task set, and an evident desire to communicate through the production of the text. It proposes either food for reflective thought by 'self', or a desire to interact with others, or both. Attention is clearly*

[1] See Csikszentmihalyi (1990), *op. cit.*

	awareness, the expression of anomie and alienation. The communicative activity is undertaken through volition and integrates self with other in a balanced way, and as an end in itself, with its own rewards. Attention is focussed on the activity in itself, and not in an 'exotelic' manner, on predicted consequences of such activity.	*signs of contextualisation within the framework of the language and the society of the speakers of that language. The production communicates either anomie, or alienation from the language, society and the task set, as if attention has been focussed on an undefined elsewhere. The author's motivation and goals lie outside the realm of the task proposed.*	*reflective thought by 'self', or much desire to interact with others. Attention appears at times distracted or confused. The text creates little sense of autonomy on the part of the 'self' that has composed it, and shows few signs of contextualisation within the framework of the language and the society of the speakers of that language. The production communicates weak volition to engage with the language, society and the task set, as if the author would prefer to focus attention on something else. The author's deeper motivation and goals lie outside the realm of the task proposed.*	*interact with others. Attention appears for the most part, focussed on the task in hand. The text creates a sense of autonomy on the part of the 'self' that has composed it, and shows recognition, either implicit or explicit, of the context of the framework of the language and the society of the speakers of that language. The production communicates a clear volition to engage with the language, society and the task set, as if the author intends to communicate. Any deeper motivation or ulterior goal appears irrelevant to the function of the production.*	*focussed on the task in hand. The text creates a sense of autonomy on the part of the 'self' that has composed it, and shows willing recognition, either implicit or explicit, of the context of the framework of the language and the society of the speakers of that language. An attempt is made to integrate form, function and content in a satisfying way. The production communicates a clear volition to engage with the language, society and the task set, as if the author has 'something significant to say'. Any deeper motivation or ulterior goal appears irrelevant to the function of the production.*	*focussed on the task in hand, and the activity gives the impression of having been pleasurable in itself. The text creates a strong sense of autonomy on the part of the 'self' that has composed it, and shows willing recognition, either implicit or explicit, of the context of the framework of the language and the society of the speakers of that language. The integration of form, function and content in a satisfying way appears to have been a goal of the author. The production communicates positive volition to engage with the language, society and the task set, as if the author has 'something significant to say'. Any deeper motivation or ulterior goal would seem completely irrelevant to the function of the production.*

APPENDIX 4

TOPICS CHOSEN FOR PRESENTATION IN THE *INTERNAL ASSESSMENT* COMPONENT FOR THE MAY 2001,2002 AND 2003 EXAMINATION SESSIONS

Internal Assessment Topics for the May 2001, 2002 and 2003 Examination Sessions

The following have been approximately categorised by the researcher in groups corresponding to the rubrics of the programme, as the *Exploration of Change,* the *Exploration of Groups*, and the *Exploration of Leisure*. It should be noted that no classification is indicated for this purpose by the candidates, *Internal Assessors,* or the examining centres.

'Exploration of Change'

May 2001 Session

- The Election of US Presidents by the Electoral College System;
- Electric Automobiles: an environmental perspective;
- The Development of Women's Rights in America and France;
- Personal Experiences of International Life;
- Marriage, Divorce and the Effects on Family Life;
- Celebrity and its Effects on Individuals;
- The Problems of the Environment;
- The Genetic Modification of Foodstuffs;

May 2002 Session

- Problems of Pollution in the Third World;
- Techniques and ethics of Biotechnological Developments;
- The Question of Cloning Human Beings;
- The Introduction of the Euro: changes in perceptions of identity;
- The Environment and Global Warming;
- The Family: yesterday and today;
- The Ethics of Genetic Science;
- My Personal Future in the World of Work;

May 2003 Session

- Budget Reforms in the French Education Ministry;
- The Future: my personal future and that of the USA;
- Environmental Problems;
- "Changes I would make to my school: security cameras";
- A Musical Revolution: Stravinsky in Paris.

'Exploration of Groups'

May 2001 Session

- Animal Rights;
- The French National Assembly: history, structure and procedures;
- Professionalism in Sport: the Situation in Canadian Ice Hockey;
- The Israel – Palestine Conflict;
- The Plight of the Homeless;
- Ethics and Morality in Philosophy;
- Violence in American Society: mass murders in schools;
- The Effects of Video-Watching;
- The Effects of Music on the Representation of Violence;
- The Status of Women in France;
- Terrorism and Justice: the Death Penalty;
- Family Life: a feminist perspective;
- My Friend Kevin: a story of drinking and driving;
- Human Relationships with Animals: an educational perspective;
- French *Châteaux*: inhabitants and their history;

- Violence in Schools in the US; A comparison of American and French law in their effects on adolescent life; Family relationships: psychological and problematic aspects;
- A comparison of American and French Education Systems (*two examples*);
- The Experience of an International Political Conference for Students held in Washington DC;
- The Use of Drugs in Sport (*two examples*);
- Violence in Commonly-Accessed Media Representations (literature, radio and television);
- Art Education in France;

May 2002 Session

- Censorship in the Art World;
- Evening Pastimes in France;
- Terrorism: the Example of Corsica;
- Terrorism: the Afghanistan question;
- Differences in political perspectives between the EU, France, the US and Canada;
- French Cinema and Theatre;
- The Significance of Muslim History;
- A Statistical Comparison of French and US Cinema Attendance and Film Preference;
- Traditional Festivities in France and French Canada;
- The French Presidential Elections: Chirac and Jospin;
- Youth Camps and Voluntary Service;
- The effects of Salaried Sport on Sporting Life;
- Corsican Current Events;
- A Concert of the Wichita Symphony Orchestra
- Traditions and Social Aspects of Cinema Going;
- The Death Penalty in France and the USA
- Encouraging or Discouraging Students in the Learning Process (*two examples provided from a single centre*);
- Comparisons between the American and the French Educational Worlds;
- Differing Family Groupings: a Miscellany.

May 2003 Session

- 'Black' Art;
- The Death Penalty: comparisons between France and the USA (*two examples*);
- The Problem of Teenage Pregnancies;
- The 14th. July Festivities in France;
- The French Co-habitation Laws;
- Tobacco Addiction;
- The Laws on Drinking Ages for Alcohol;
- My Family: my sister – portrait of a loved one;
- My Family: my mother – portrait of a loved one ;
- Mass Media;
- The Word 'Love', according to the Bible;
- Being Young Today;
- My Family and the War in Iraq;
- The Ivory Coast;
- History of the French Resistance: Gérard Chauvet and the Second World War;
- '*L'Etranger*' by Albert Camus (*three examples from the same centre*);
- Immigration in France;
- Prostitution in French Law and Practice;
- Terrorism;
- American Perspectives on Life in Luxembourg;
- Me, My Mum, My School and Judith Resnick, astronaut from the 'Challenger' Space Shuttle Disaster.

'Exploration of the World of Leisure'

May 2001 Session

- The Origins of the Rock Group: *Public Enemy*;
- Internet Music Services: the *Napster* case;
- Ballet;
- The Sport of Tieball;
- Leisure Activities in General: a comparison between adolescent life in America and France;
- American and French preferences in cinema-going;
- Canadian Ice Hockey: the State of Play;

May 2002 Session

- Injustices in the World of Sport;
- Objections to Female Wrestling;
- Popular Music Journalism: comparisons between Britain, Germany and the US;
- Animated Cartoons in France and the US;
- Sport in France;
- My Preferred Films;
- Josephine Baker: Dancer and Singer;
- The French Electronic Music Group: AIR;
- Tourism: Some Advantages and Disadvantages;
- The Use of Leisure Time in the United States.

May 2003 Session

- Leisure Time in the USA;
- French Cartoons: Astérix;
- The Importance of Sport and the Advantages it brings;
- CISV (International Holiday Village Centres);
- International Football;
- My Holiday in France;
- French Cinema: latest releases ;
- The Film "My Greek Wedding";
- Two Racing Greyhound Bitches that I adopted;

Topics that are ambiguous for categorisation

May 2001 Session

- The Experience of Bilingualism;
- Psychology as an Academic Discipline;
- AIDS (*two examples, one general, one limited to the situation in Canada*);
- Suicide;
- Genetic Science: foetal research (*two examples*);
- Health Issues: alcoholism and drug- taking (*three examples*);
- The Cold War: the issue of nuclear weaponry;
- Mad Cow Disease;

- Michelangelo and a School Trip to Florence and Rome;
- Robotics and the Design and Construction of Robots;

<u>May 2002 Session</u>

- A Personal Tragi-Romance: January 2002.

<u>May 2003 Session</u>

- Psychology and Art;
- Archaeology in Iraq;
- The Maginot Line;
- Victor Hugo : the Man and his Work;
- The War in Iraq (*four examples, three from a single centre*);
- Sleep: dreams and reality;
- Art in The Louvre;
- Fairy Tales by Charles Perrault;
- Alzheimer's Disease.

APPENDIX 5

THE COMPARISON OF SCORE AND GRADE DIFFERENCES FOR 100 CANDIDATES FOR *INTERNAL ASSESSMENT*: May 2001 and May 2002

Table 6.11

Comparison of Score and Grade Differences for 100 Individual Candidates for *Internal Assessment*, May 2001 and May 2002
(with anomalous cases of incomplete data eliminated)

Candidate	Mark with IBO scheme	Mark with Van Lier scheme	Difference in mark	Difference in IBO grading
1	28	20.5	-7.5	- 2
2	8	4.5	-3.5	- 1
3	9	5.5	-3.5	- 1
4	8	5.5	-2.5	- 1
5	28	25.5	-2.5	- 1
6	27	25	-2	- 1
7	10	10	0	0
8	13	13	0	0
9	14	14	0	0
10	14	14	0	0
11	14	14	0	0
12	15	15	0	0
13	15	15	0	0
14	16	16	0	0
15	16	16	0	0
16	16	16	0	0
17	20	20	0	0
18	20	20	0	0
19	20	20	0	0
20	12	12.5	0.5	0

Candidate	Mark with IBO scheme	Mark with Van Lier scheme	Difference in mark	Difference in IBO grading
21	12	12.5	0.5	0
22	13	13.5	0.5	0
23	15	15.5	0.5	0
24	20	20.5	0.5	0
25	20	20.5	0.5	0
26	26	26.5	0.5	0
27	10	11	1	0
28	10	11	1	0
29	10	11	1	0
30	11	12	1	0
31	11	12	1	0
32	13	14	1	1
33	13	14	1	1
34	14	15	1	0
35	15	16	1	0
36	15	16	1	0
37	16	17	1	0
38	17	18	1	1
39	17	18	1	1
40	17	18	1	1
41	17	18	1	1
42	18	19	1	0
43	18	19	1	0
44	18	19	1	0
45	19	20	1	0
46	19	20	1	0
47	19	20	1	0
48	19	20	1	0
49	19	20	1	0
50	20	21	1	0
51	9	10.5	1.5	0
52	9	10.5	1.5	0
53	16	17.5	1.5	0
54	17	18.5	1.5	1
55	17	18.5	1.5	1
56	17	18.5	1.5	1
57	21	22.5	1.5	0

Candidate	Mark with IBO scheme	Mark with Van Lier scheme	Difference in mark	Difference in IBO grading
58	18	20	2	0
59	18	20	2	0
60	18	20	2	0
61	18	20	2	0
62	20	22	2	0
63	20	22	2	0
64	20	22	2	0
65	20	22	2	0
66	21	23	2	0
67	21	23	2	0
68	22	24	2	1
69	22	24	2	1
70	21	23.5	2.5	0
71	21	23.5	2.5	0
72	22	24.5	2.5	1
73	23	25.5	2.5	1
74	15	18	3	1
75	22	25	3	1
76	22	25	3	1
77	22	25	3	1
78	25	28	3	1
79	25	28	3	1
80	25	28	3	1
81	25	28	3	1
82	26	29	3	1
83	26	29	3	1
84	26	29	3	1
85	27	30	3	0
86	27	30	3	0
87	7	10.5	3.5	1
88	25	28.5	3.5	1
89	25	28.5	3.5	1
90	24	28	4	1
91	24	28	4	1
92	25	29	4	1
93	26	30	4	1
94	12	16.5	4.5	1

Candidate	Mark with IBO scheme	Mark with Van Lier scheme	Difference in mark	Difference in IBO grading
95	23	27	5	2
96	23	27	5	2
97	23	27	5	2
98	20	25.5	5.5	1
99	23	27.5	5.5	2
100	23	27.5	5.5	2